Contributions to the theoretical analysis of the algorithms with adversarial and dependent data

Oleksandr Zadorozhnyi

Institut für Mathematik
Mathematische Statistik

Contributions to the theoretical analysis of the algorithms with adversarial and dependent data

Dissertation (version)
zur Erlangung des akademischen Grades
"doctor rerum naturalium"
(Dr. rer. nat.)
in der Wissenschaftsdisziplin "Stochastik"

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
der Universität Potsdam

von
Oleksandr Zadorozhnyi

Datum/Ort der Disputation: 08.07.2021/Potsdam

Hauptbetreuer: Prof. Dr. Gilles Blanchard
Betreuer: Prof. Dr. Tobias Scheffer
Gutachter: Prof.Dr. Aurélien Garivier
Gutachter: Prof. Dr. Ingo Steinwart

Disclaimer: The text below is strictly speaking neither the sub- nor the superset of the text of mine dissertation which was submitted in March 2021 and defended in July 2021 at the University of Potsdam. In particular, it contains some corrected typos and minor refinements/changes in formulations. A printed version of the submitted dissertation one can find in the library of Uni Potsdam.

Acknowledgements

I want to express my sincere gratitude to my scientific supervisor Gilles Blanchard for his time, dedication and support which was really very important for me during these years. I want to thank to Tobias Scheffer and his group at the Institute of Informatics with whom I had a good chance to get acquainted with many interesting applied topics in machine learning and for being friendly. I want to thank to my coauthors and advisors Alexandra Carpentier, Pierre Gaillard, Sebastién Gerchinovitz, Alessandro Rudi who (as well as Gilles Blanchard) stayed behind most of the ideas and contributed very much to our joint works and also were always supportive. I want also to thank reviewers Aurélien Garivier and Ingo Steinwart for agreeing to review this thesis and for their time. I acknowledge SFB 1294 and CDFA 01-18 for financial and travel support during the years of my doctoral studies.

Allgemeinverständliche Zusammenfassung auf Deutsch

In dieser Arbeit beschäftigen wir uns mit den theoretischen Aspekten vom Maschienellen Lernen. In diesem Gebiet beschäftigen sich Leute mit den folgenden Problemen: wie man anhand von Daten die Phenomene in der Natur und im alltäglichen Leben beschreibt und anhand von Daten die Strategien für präzisere Vorhersagen entwickelt. Ein typisches Beispiel wäre die Vorhersage der Temperatur in einem Gebiet anhand von vorhandenen Messungen. Ein anderes Beispiel wäre die Objekterkennung, bzw. Objektklassifikation von neuen Objekten anhand von gegebenen Bildern. In solchen Beispielen die Daten können auch komplexere Struktur haben oder von einem komplizierterem Prozess mit Gedächtnis generiert sein. Im Fokus dieser Arbeit sind die theoretischen Aspekte von Algorithmen des Maschinelles Lernen die anhand den abhängigen oder arbitrarän Daten konstruiert sind. In einem Kapitel dieser Arbeit untersuchen wir die sogenannate Regularisierungmethoden des Statistischen Lernens mit den Werten in hochdimensionalen Räumen, so dass man zwischen beliebigen Elementen dieser Räumen Distanz (Metric) definieren kann. Diese sind die bekannte Methoden, die im Bereich von Lernen aus unabhängigen Beispielen vorher anwegendet worden sind und in dieser Arbeit unter der Perspective von Lernen aus Daten mit Abhängigkeit anylisiert sind. Um diese Aufgabe zu verwollständigen, entwickeln wir dafür das notwendige technische Toolbox, die Konzentrationsungleichungen, die die Kontorolle der Fluktuationen von zufalligen Elementen ermöglichen. Die Neuheit der Ergebnissen liegt daran, dass man anstatt der stochastischen Unabhängigkeit der Daten die sogennante Annahme der schwachen Abhängigkeit über die Verteilung der Daten betrachtet. Man beobachtet dabei einen interessanten Effekt, nämlich, dass in manchen Szenarien das Lernen aus abhängigen Daten genauso kompliziert (im Sinne von theoretischen Komplexität) ist als das Lernen aus unabhängigen Daten.

In einem anderen grös teil der Arbeit betrachten wir das Szenario vom Sequentiellen Lernen. In diesem Framework das Lernen aus dem Daten kann als ein Spiel zwischen dem Lerner und Gegner (auch "Adversary" gennant) modelliert werden. Der Hauptunterschied zu dem statistischen Lernen von Beispielen liegt darin, dass der Lerner die Vorherage (das "Forecaster") aktualisiert sobald das neue Beispiel vorhanden ist. Der Ziel dabei einen online Forecaster zu entwickeln, der vergleichbar gut zu den besten offline gefundenen Forecaster ist. Die Stärke des Algorithmus ist nun gemessen als die Differenz zwischen dem Verlust der Sequenz von Vorhersagen und dem Verlust, die man anhand von einem belibigem konstanten Forecaster aus dem gegebenen Klass von Forecaster erziehlen kann.

Weiterhin, im Framework von Sequentiellen Lernen beschäftigen wir uns mit dem sogennanten "Multi-Armed-Bandit" Problem. Dieses Modell kann mithilfe von einem Experiment mit der Medikationsuntersuchung modelliert werden, in dem jedem der ankommenden Patienten einen (und nur einen) der vorhadenen Medikamenten vorgeschrieben wird und die Effiezienz ("Reward") der Behandlung gemessen wird. Das Ziel dabei so viel wie möglich Patienten zu heilen. Das schwieriege an dem Problem ist, dass bei einer Untersuchung nur die Effiezienz von einem der Medikamenten untersucht werden konnte (und nicht von alle Menge der Elementen). Das Beispiel fällt unter dem Hut von Sequentielles Lernen mit dem partiellen (oder sogennanten 'Bandit' feedback). In dieser Arbeit untersuchen wir das Szenario indem die vorhandene stochastiche rewards Abhängigkeitsstruktur (auch von schwach-abhängiger Natur) besitzen.

In dem letzten Kapitel untersuchen wir ein Problem der Konzentrationsungleichungen für zufällige Feldern in $\mathbb{N}^d$. Dieses Problem ist erstmal von Bedeutung, da die alle bekannte Ergebnisse auf die unterliegende Klasse von schwach-abhängigen realwertigen zufälligen Prozessen nicht zu den optimalen Raten für die Abweichungen der partiellen Sumenn (entweder in Wahrscheinlichkeit oder in integral norm) führen. Weiterhin, die Ergebnisse können auch im Statistischen Online Lernen Setting angewendet und zu Banach-wertigen Zufallsvariablen erweitert werden. Das Zusammenführen von bekannten Techniken aus Statistik (Chaining method) sowie aus Wahrscheinlichkeitstheorie (Martingale-difference approach) führt dabei zu den neuen Methode die auch weiter erforscht werden kann und zu dem tieferen Verständnis von Effekten der schwach-abhängigen Daten auf das Verhalten von statistichen Risiko von breitem Spektrum von Methoden führen kann.

Contents

Chapter 1

Introduction

Contents

1.1 Machine learning framework and statistical learning

1.1.1 Introduction to the problems of machine learning.

Machine learning is a rapidly developing branch of science which includes applications in many fields. It is already a well-established part of industry and continues to become part of day-to-day life. It studies the principles, methods and techniques of devising data-dependent decision rules which enable intelligent-like decisions.

An important aspect of developing decision rules is being able to provide theoretical guarantees for the predictions given by an algorithm, and thus, to be able to learn from data. There are different specific frameworks within the setting of learning from data observations. For example, one framework is the supervised learning in which a pair of covariate-label (x_t, y_t) is produced by the environment, covariate x_t is revealed, and the learner aims to predict a hidden label y_t based on the information contained in the covariate x_t and all the information which was available before. Another framework is the unsupervised learning where a learner deals with the set of covariates and the typical task is to discover the inherent grouping between the sets of available objects. This setting is also named clustering. A different framework is reinforcement learning where the goal is to learn the dynamic structure of the environment by interacting with it in a sequential fashion and by obtaining rewards for taking actions which can influence the dynamics of the environment.

In this work we mainly focus on the particular aspect of an another major setting: learning from a given (batch) data sample under certain dependency assumption and sequential (online) learning. In the batch framework, a learner has access to a set of data $\mathcal{D}$ and devises some decision rules based on this sample. In the sequential setting, the decision rule is devised (or updated) each time a new observation of the data sample is arrived from the data-stream, and in certain cases it influences the procedure of how the new output is generated. On a theoretical level, a difference between the settings of online and batch learning was firstly highlighted in the work of Thompson (1933), where the question of devising statistically efficient procedures which distinguish between the effects of two medications based on as few observations as possible was raised. In this setting, the efficiency is based on the average "goodness" of patients' treatments. A typical example of the problem which can be considered in both frameworks is the mail-classification task. In this task a learner (which in this case may be a spam-filter or some intelligent computer system) either possesses a database of emails or obtains mails in a stream fashion. Given the text of observed emails, and possibly some additional mail-attributes (like domain names, time when the email was sent, etc.), a learner aims to devise a decision rule which for a new mail can classify it as being "spam" or "not spam". This is a typical example of a binary classification problem which can be generalized for the classification of a larger (but finite) number of classes. A typical example of multi-class classification problem is the image-recognition problem, where the database consists of a large number of images, each with a label that characterizes what is actually depicted on the image. Similarly as in the case of binary classification the goal is to devise a data-dependent decision rule which, given a new image (typically represented as a multidimensional matrix with each matrix entry being a $3-d$ vector of intensity measurements on the rgb levels), outputs one of the many available categories. A standard example of a regression problem is the prediction of housing prices, or the rental price modelling for the dwelling given information about its location, area, and history of previous changes of price or rent. In this case, the database is represented as a collection of data which describes an object with a given label - price of the house or cost of rent. For a new unlabeled item in the database, a decision rule aims to predict its label, that is the price of the house. Other usage of regression-based machine learning algorithms can be found in banking, market profit prediction (prediction of the profit of a customer of a marketing company Schmidt (2019)), and prediction of the forest areas which will be potentially burned by forest fires (Cortez and Morais (2007)). In most of these cases, the data can be corrupted by noise which typically has some dynamic.

Notice that in all these concepts the "learning" is not defined rigorously. Informally one can speak of "learning" as of devising decision rules based on the observations of the environment. Several attempts have been made to provide proper theoretical foundations and give mathematical justification for the performance of learning algorithms both in the frameworks of batch learning and online learning (see ex. Steinwart and Christmann (2008), Shalev-Shwartz (2007) and Cesa-Bianchi (1999a),Smola and Schölkopf (2002),Bishop (2006)). We first consider the setting of (batch) learning from examples. In this setting it is assumed that the data sample of n of pairs $\mathcal{D}_n = (x_s, y_s)_{s=1}^n \in (\mathcal{X} \times \mathcal{Y})$, where $\mathcal{X}$ is some vector space and $\mathcal{Y}$ is subset of real line is available from the beginning and the task is to construct a data-dependent estimator given x for an unknown (new) label y. The latter pair is assumed to have distribution $\nu(x, y)$ over $\mathcal{X} \times \mathcal{Y}$ while data sample $\mathcal{D}$ is supposed to be generated from the $n-$step trajectory of some stochastic process $(Z_t = (X_t, Y_t))_{t\geq 1}$. In Chapter 2 of this thesis we consider the framework of statistical learning from dependent data observations.

1.1.2 Mathematical aspects of statistical learning theory.

In this section we introduce some definitions and probabilistic notations which are standard in the framework of statistical learning. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be some probability space. Denote $L_p(\Omega, \mathcal{F}, \mathbb{P})$, $1 \leq p < \infty$ to be the standard space of equivalence classes with respect to measure $\mathbb{P}$ of real $p-$integrable functions equipped with the norm $\|f\|_p = \|f\|_{L_p(\Omega,\mathcal{A},\mathbb{P})} := \int_\Omega |f|^p(\omega)\mathbb{P}(d\omega)$ and denote for $[f]_{\mathbb{P}} \in L_p(\Omega, \mathcal{F}, \mathbb{P})$

the $\mathbb{P}-$equivalence class of any map $f : \mathcal{X} \mapsto \mathbb{R}$ such that $\int_\Omega |f|^p(\omega)\mathbb{P}(d\omega) < \infty$. Let $L_\infty(\Omega, \mathcal{F}, \mathbb{P})$ be the set of essentially bounded real functions, namely such that $\|f\|_\infty := \operatorname{ess\,sup}_\omega |f(\omega)| = \inf\{c \in \mathbb{R}_+ : \mathbb{P}[|f(\omega)| \geq c] = 0\} < \infty$. In the case when $\mathbb{P}$ is Lebesgue measure over $\mathcal{X}$ we use the shorthand notation $L_p(\mathcal{X})$. Lastly for a Banach space $(\mathcal{W}, \|\cdot\|_\mathcal{W})$ we denote $B_\mathcal{W}(R), \overline{B}_\mathcal{W}(R)$ to be the open and the closed ball of radius R centered in origin of $\mathcal{W}$.

The classical supervised learning scenario is a problem of learning a function from (random) examples, and in this framework it can be formulated as follows. Let $(\mathcal{X}, \mathcal{Y})$ be a pair of vector spaces where $\mathcal{X}$ is assumed to be a normed Polish space, and $\mathcal{Y}$ is assumed to be a closed subset of $\mathbb{R}$. $\mathcal{X}$ is called the input space and $\mathcal{Y}$ is called the output space. Define $\mathcal{B}(\mathcal{X})$, $\mathcal{B}(\mathcal{Y})$ to be Borel $\sigma-$algebras of open sets over $\mathcal{X}, \mathcal{Y}$, let ν be some probability measure over $((\mathcal{X} \times \mathcal{Y}), \mathcal{B}(\mathcal{X} \times \mathcal{Y}))$. Consider a $\mathcal{X} \times \mathcal{Y}$-valued stationary process $(Z_t = (X_t, Y_t))_{t\in\mathbb{N}}$ such that $Z_t \sim \nu$ and distribution of $(Z_t)_{t\in\mathbb{N}}$ is defined through its canonical version over a probability space $\Big((\mathcal{X} \times \mathcal{Y})^\mathbb{N}, \mathcal{B}\Big((\mathcal{X} \times \mathcal{Y})^\mathbb{N}\Big), \mathbb{P}\Big)$. In the framework of learning from examples it is always assumed that there is some intristic dependence between the outputs y_t and inputs x_t. We write μ for the the $X-$marginal of ν, and $\nu(\cdot|x)$ for the conditional distribution of Y_t over $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ given $x_t = x$. Notice that probability measure $\mathbb{P}$ is known to the learner only through the sample of size n, $\mathcal{D}_n = (x_t, y_t)_{t=1}^n$. In the simplest case where $\mathbb{P}$ is a product measure, we have that pairs $Z_t = (X_t, Y_t)_{t\in\mathbb{N}}$ is the realization of the sequence of i.i.d. random variables $(X_t, Y_t)_{t=1}^n$, such that $(X_t, Y_t) \sim \nu(x, y)$. In the framework of learning from examples, the goal of the learner is to construct ("learn") a prediction rule $f_{\mathcal{D}_n} : \mathcal{X} \mapsto \mathbb{R}$ to predict label y in the forthcoming pair (x, y) based on the sample $\mathcal{D}_n = \{x_t, y_t\}_{t=1}^n$. Notice that we don't restrict of the value $f_\mathcal{D}(x)$ to the prediction set $\mathcal{Y}$, however this can be done by considering the so-called clipping procedure (see more details in Steinwart and Christmann (2008)). We define the following objects which are necessary for smoothness of further narrative.

Definition 1.1.1. We call a sequence of maps $\mathcal{L} = (\mathcal{L}_n)_{n\geq 1}$ a *learning method* if for any data-sample $\mathcal{D}_n = (x_t, y_t)_{t=1}^n \in (\mathcal{X} \times \mathcal{Y})^n, n \geq 1$, $\mathcal{L}_n$ associates a prediction rule $f_{\mathcal{D}_n} \in \mathbb{R}^\mathcal{X}$. We say that learning method $\mathcal{L}$ is *measurable* if the map $\mathcal{L}_n : (\mathcal{X} \times \mathcal{Y})^n \times \mathcal{X} \mapsto \mathbb{R}$, $(\mathcal{D}_n, x) \mapsto f_{\mathcal{D}_n}(x)$, is measurable with respect to the completion of the product $\sigma-$algebra on $(\mathcal{X} \times \mathcal{Y})^n \times \mathcal{X}$, for every $n \in \mathbb{N}$.

To quantify the "goodness" of the prediction $f(x)$ for a label y, we define the concept of loss function (or simply loss).

Definition 1.1.2. We say that loss L is any measurable function $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \mapsto \mathbb{R}_+$.

In statistical learning theory, for a given loss-function $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \mapsto \mathbb{R}$ and unknown distribution ν, one measures the "quality" of the prediction of a decision rule $f : \mathcal{X} \mapsto \mathbb{R}$ by means of the expected risk, which is defined as follows.

Definition 1.1.3. For a loss function $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \mapsto \mathbb{R}_+$ and measure ν over $\mathcal{X} \times \mathcal{Y}$, the *expected risk* of a measurable function $f : \mathcal{X} \mapsto \mathbb{R}$ is defined as:

$$R_{L,\nu}(f) = \mathbb{E}_\nu[L(X, Y, f(X))] = \int_\mathcal{X} \int_\mathcal{Y} L(x, y, f(x)) d\nu(y|x) d\mu(x).$$

For a given data-sample $\mathcal{D}_n = (x_t, y_t)_{t=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$, we denote $\mathbb{P}_n := \frac{1}{n}\sum_{i=1}^n \delta(X_i, Y_i)$ to be the empirical measure associated with $\mathcal{D}_n$. We consider the *empirical risk* of decision rule f as a risk with respect to empirical measure $\mathbb{P}_n$ and define

$$R_{L,\mathbb{P}_n}(f) := \frac{1}{n}\sum_{t=1}^n L(x_t, y_t, f(x_t)).$$

For a decision rule $f_{\mathcal{D}_n}$, the quantity $R_{L,\nu}(f_{\mathcal{D}_n})$ depends on the distribution of the data sample $\mathcal{D}_n$. Furthermore, the loss of decision rule $f_{\mathcal{D}_n}$ is measured as expected risk with respect to the measure ν

as if (x, y) where stochastically independent of the sample $\mathcal{D}_n$. This approach is different (while being widely used in recent works, see for example Yu (1994), Meir (2000), Vidyasagar (2003) and Lozano et al. (2005)) to the situation when evaluating against $\mathbb{Q}_{n+1}(\cdot|\mathcal{D}_n)$ (conditional distribution of a next pair (x_{n+1}, y_{n+1}) given the data-sample $\mathcal{D}_n$ under the restriction of the measure $\mathbb{P}$ to a finite dimensional distribution of $\{X_t, Y_t\}_{t=1}^{n+1}$). The latter can be proven to exist by disintegration, see also discussion in Remark 1.1.6). This usage can be further justified from the ergodic perspective. Namely, in the setting in which the learning method $\mathcal{L}$ outputs the prediction rule $f_{\mathcal{D}_n}$ based on the sample $\mathcal{D}_n$, a learner aims to apply it to predict not only the label y_{n+1} but also labels in all forthcoming pairs $(x_s, y_s)_{s>n}$. If $(Z_t)_{t\geq 1}$ is a stationary stochastic process as above and loss function ℓ is bounded, by the Ergodic Theorem (Theorem 20.14 in Klenke (2010)) the time-averaged loss $\lim_{m\to\infty} \frac{1}{m}\sum_{j=1}^{m} L(x_{n+j}, y_{n+j}, \widehat{f}_{\mathcal{D}_n}(x_{n+j}))$ is $\mathbb{P}-$almost surely equal to the $R_{L,\nu}(f_{\mathcal{D}_n})$. That is $R_{L,\nu}(f_{\mathcal{D}_n})$ is a good proxy for the risk we obtain by applying a decision rule $f_{\mathcal{D}_n}$ learned over the n samples when $n \mapsto \infty$.

Definition 1.1.4. For a given loss function L and distribution ν we define the smallest risk a measurable function can achieve as follows:

$$R^*_{L,\nu} = \inf_{f:\mathcal{X}\mapsto\mathbb{R}} R_{L,\nu}(f). \tag{1.1}$$

We refer to $R^*_{L,\nu}$ as to the *Bayes risk* (with respect to measure ν and loss L). In addition, any measurable function f^* such that $R_{L,\nu}(f) = R^*_{L,\nu}$ is called a *Bayes decision function*.

In the setting of statistical learning, the distribution $\mathbb{P}$ is available to the learner only through the sample $\mathcal{D}_n = (X_t, Y_t)_{t=1}^n$ from the $n-$dimensional restriction of the measure $\mathbb{P}$. One consider the performance of decision rule $f_{\mathcal{D}_n}$ given by some learning algorithm $\mathcal{L}$ by considering its expected risk $R_{L,\nu}(f_{\mathcal{D}_n})$ in comparison to $R_{L,\nu}(f^*)$. A fundamental question of learning theory is whether the risk of the empirical decision rule $f_{\mathcal{D}_n}$ converges (in probability or in expectation) to the risk minimizer f^*. If so, one wants to quantify the convergence $R_{L,\nu}(f_{\mathcal{D}_n})$ to $R^*_{L,\nu}$ as $n \to \infty$ and to have explicit bounds on the quantity $\mathbb{P}\Big(R_{L,\nu}(f_{\mathcal{D}_n}) - R^*_{L,\nu} \geq \varepsilon\Big)$ for any $\varepsilon > 0$ (either in the asymptotic regime or for fixed n), or to control the deviation in expectation, i.e. $\mathbb{E}\Big[R_{L,\nu}(f_{\mathcal{D}_n}) - R^*_{L,\nu}\Big]$. We give the following standard definition of the algorithm for which the convergence in probability occurs.

Definition 1.1.5. The algorithm $\mathcal{L} = (\mathcal{L}_n)_{n\geq 0}$, $\mathcal{L}_n : \mathcal{D}_n \mapsto \mathcal{Y}^{\mathcal{X}}$ which produces a measurable decision rule $f_{\mathcal{D}_n} : \mathcal{X} \mapsto \mathbb{R}$ based on a data sample $\mathcal{D}_n$ of size n is called consistent under distribution $\mathbb{P}$ if for all $\varepsilon > 0$ $\lim_{n\to\infty} \mathbb{P}\big(R_{L,\nu}(f_{\mathcal{D}_n}) - R^*_{L,\nu} \geq \varepsilon\big) = 0$. It is called strongly consistent if $R_{L,\nu}(f_{\mathcal{D}_n})$ converges to $R^*_{L,\mu}$ $\mathbb{P}-$almost surely. Similarly, one says that $\mathcal{L}$ is consistent in expectation if

$$\lim_{n\mapsto\infty} \mathbb{E}\big[|R_{L,\nu}(f_{\mathcal{D}_n}) - R^*_{L,\nu}|\big] = 0$$

Remark 1.1.6. Notice that in the latter definition of consistency we always consider the quantities under original distribution $\mathbb{P}$ of stochastic process $(Z_t)_{t\in\mathbb{N}}$ while distribution of $R_{L,\nu}(f_{\mathcal{D}_n}) - R^*_{L,\nu}$ depends only on the $(Z_t = (X_t, Y_t))_{t=1}^n$. In the case of of i.i.d. process $\mathbb{P}$ is a product measure, thus one can write $\mathbb{P}\Big(R_{L,\nu}(f_{\mathcal{D}_n}) - R^*_{L,\nu} \geq \varepsilon\Big) = \mathbb{Q}_n\Big(R_{L,\nu}(f_{\mathcal{D}_n}) - R^*_{L,\nu} \geq \varepsilon\Big)$, where $\mathbb{Q}_n = \nu^{\otimes n}$. In the case of general stochastic process with the given measure $\mathbb{P}$, for every n we can take $\mathbb{Q}_n$ such that $\mathbb{Q}_n$ is obtained as disintegration of the measure $\mathbb{P}$ over measure $\mathbb{Q}_n$ over $((\mathcal{X}\times\mathcal{Y})^n, \mathcal{B}((\mathcal{X}\times\mathcal{Y})^n))$ and stochastic kernel $\widehat{k}_n : ((\mathcal{X}\times\mathcal{Y})^n, \mathcal{B}((\mathcal{X}\times\mathcal{Y})^n)) \mapsto \Big((\mathcal{X}\times\mathcal{Y})^{\mathbb{N}}, \mathcal{B}\Big((\mathcal{X}\times\mathcal{Y})^{\mathbb{N}}\Big)\Big)$, i.e. $\mathbb{P} = \mathbb{Q}_n \otimes \widehat{k}_n$ (see Theorem 1.23 in Kallenberg (2017) to ensure the existence of such $\mathbb{Q}_n$ and moreover for the proof that we can take $\mathbb{Q}(\cdot) = \mathbb{P}\big(\cdot \times \Omega^{\mathbb{N}}\big)$). By Fubini's Theorem (see Klenke (2010), Theorem 14.16 on p. 278) it is easy to check that in this case it holds $\mathbb{P}\Big(R_{L,\nu}(f_{\mathcal{D}_n}) - R^*_{L,\nu} \geq \varepsilon\Big) = \mathbb{Q}_n\Big(R_{L,\nu}(f_{\mathcal{D}_n}) - R^*_{L,\nu} \geq \varepsilon\Big)$.

It is easy to observe (by using Markov's inequality) that consistency in expectation implies consistency in probability. Notice that whether an estimator $f_{\mathcal{D}_n}$ is (strongly) consistent or not depends on the underlying distribution $\mathbb{P}$. Thus, in general, a learning algorithm which is (strongly) consistent for one distribution is not (strongly) consistent for another. Therefore, one is interested in having a decision rule which is consistent uniformly over all (or at least over some class of) distributions $\mathbb{P}$. The algorithm is called *universally consistent* if for all $\varepsilon > 0$, it holds that

$$\limsup_{n\to\infty} \mathbb{P}\big(R_{L,\nu}(f_{\mathcal{D}_n}) - R^*_{L,\nu} \geq \varepsilon\big) = \limsup_{n\to\infty} \mathbb{Q}_n\big(R_{L,\nu}(f_{\mathcal{D}_n}) - R^*_{L,\nu} \geq \varepsilon\big) = 0$$

for any distribution $\mathbb{P}$ and its sequence of $n-$dimensional restrictions $(\mathbb{Q}_n)_{n\geq 0}$. Universal consistency is proven for various methods under the i.i.d. noise assumption (see for example in Christmann and Steinwart (2007) for the learning problem with arbitrary convex loss) and with dependencies in samples (see ex. Zou et al. (2009) for consistency of empirical risk minimization procedure under uniformly ergodic Markov chain assumption on the distribution of the process). However, universal consistency is purely an asymptotic property of a learning algorithms and does not give information about how *fast* $R_{L,\nu}(f_{\mathcal{D}_n})$ converges to $R^*_{L,\nu}$. In other words, it does not characterize how well the estimator has learned data from a fixed sample of size n. The latter notion is characterized by introducing the concept of *learning rate*. We give the following definition of the learning rate (w.r.t. to a given probability measure $\mathbb{P}$).

Definition 1.1.7. For a given confidence $\tau \in [0,1]$, measure $\mathbb{P}$, the learning method $\mathcal{L} : \mathcal{D}_n \mapsto \mathcal{Y}^{\mathcal{X}}$ which outputs the decision rule $f_{\mathcal{D}_n} : \mathcal{X} \mapsto \mathbb{R}$ based on the sample $\mathcal{D}_n = \{x_t, y_t\}_{t=1}^n$ from the stochastic process $Z_t = (X_t, Y_t)_{t\geq 1}$ is said to *learn with rate* $(a_n)_{n\geq 0}$ if there exists a constant $c_\tau > 0$ and number $n_0 \in \mathbb{N}$ such that:

$$\mathbb{P}\big(R_{L,\nu}(f_{\mathcal{D}_n}) \leq R^*_{L,\nu} + c_\tau a_n\big) \geq 1 - \tau, \tag{1.2}$$

for all $n \geq n_0$. If the learning $\mathcal{L}$ method learns an optimal decision rule f^* with rate $(a_n)_{n\geq 0}$, then the sequence $(a_n)_{n\geq 0}$ is called the *learning rate* for method $\mathcal{L}$ under measure $\mathbb{P}$.

Notice that the deviation bound from Equation (1.2) and the rate depend on the underlying measure $\mathbb{P}$. Ideally, for a fixed confidence τ, one would like to have a rate $(a_n)_{n\geq 0}$ such that there exists a learning algorithm for which Equation 1.2 holds for any measure $\mathbb{P}$. However, there is no learning method which learns over all distributions $\mathbb{P}$ over $\Big((\mathcal{X}\times\mathcal{Y})^{\mathbb{N}}, \mathcal{B}(\mathcal{X}\times\mathcal{Y})^{\mathbb{N}}\Big)$. This is justified by the following result (which is called the "no-free-lunch" (see Corollary 6.8 in Steinwart and Christmann (2008) or Theorem 7.2 in Devroye et al. (1996)) Theorem which holds in expectation already in the case when $\mathbb{P}$ is a product measure.

Theorem 1.1.8 (Theorem 7.2 in Devroye et al. (1996); also Theorem 6.6. in Steinwart and Christmann (2008))**.** *Consider any decreasing sequence $(a_n)_{n\geq 0} \in [0, \frac{1}{16}]$ that converges to 0, $\mathcal{Y} = \{-1, 1\}$ and a probability space $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mu)$ such that μ is atom-free. Let $L(x, y, f) = \mathbb{I}_{f(x)\neq y}$. For every measurable learning method $\mathcal{L} : \mathcal{D}_n \mapsto \mathcal{Y}^{\mathcal{X}}$ there exists a distribution ν over $(\mathcal{X}\times\mathcal{Y}, \mathcal{B}(\mathcal{X}\times\mathcal{Y}))$ with $X-$marginal μ such that $\mathcal{R}^*_{L,\nu} = 0$ whereas $\mathbb{E}_{\nu^{\otimes n}}[R_{L,\nu}(f_{\mathcal{D}_n})] \geq a_n$ for a decision rule $f_{\mathcal{D}_n} : \mathcal{X} \mapsto \mathbb{R}$ which is returned by algorithm $\mathcal{L}$.*

The above result informally states that that there exists a measure with zero Bayes risk under zero-one loss function such that any learning method $\mathcal{L}$ may need an arbitrarily large number of observations to achieve an expected risk smaller than some given value $\varepsilon < \frac{1}{16}$, provided the given measure ν is complicated enough (and depends on the number of observations n). The only way to overcome this issue and construct learning algorithms which have certain learning rates is by introducing further assumptions on the data-generating distributions $\mathbb{P}$. From one side this seems to be like an escape from the problem,

since in practice there is almost no way to check whether underlying marginal distribution possesses the given assumption. From the other side, by establishing learning rates for some learning method $\mathcal{L}$ under different assumptions on measure $\mathbb{P}$, one can characterize for which distributions some given learning method learns faster. This may be of interest for a practitioner who, based on the domain knowledge, has to decide which method would be preferable to use for certain applications. To describe the setting of convergence of the decision rule to the Bayes decision rule precisely, we assume that $\mathbb{P}$ belongs to some distribution class $\mathcal{P}$ of measures over $\left(\mathcal{Z}^{\mathbb{N}}, \mathcal{B}\left(\mathcal{Z}^{\mathbb{N}}\right)\right)$ where we denoted $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$.

We refer to the class $\mathcal{P}$ as the prior class. Taking inspiration from the works Caponetto and De Vito (2005) and Blanchard and Mücke (2018) we define the following concepts which characterize the convergence rates of learning algorithms over the given prior class $\mathcal{P}$.

Definition 1.1.9. A non-increasing sequence $(a_n)_{n\geq 1}$ is called the *strong upper rate* of convergence in probability of a learning algorithm $\mathcal{L}$ over the prior class $\mathcal{P}$ if, for the decision rule $f_{\mathcal{D}_n}$ returned by the algorithm $\mathcal{L}$ based on the sample $\mathcal{D}_n$, it holds that:

$$\lim_{\tau\to\infty} \limsup_{n\mapsto\infty} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\left(R_{L,\nu}(f_{\mathcal{D}_n}) - R^*_{L,\nu} > \tau a_n\right) = 0. \tag{1.3}$$

The sequence $(a_n)_{n\geq 1}$ is called the lower *minimax rate of convergence* over the prior class $\mathcal{P}$ if

$$\liminf_{n\mapsto\infty} \inf_{\mathcal{L}_n} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\left(R_{L,\nu}\left(f_{\mathcal{D}_n}\right) - R^*_{L,\nu} > a_n\right) > 0, \tag{1.4}$$

where the infimum is taken over all measurable maps $\mathcal{L}_n : \mathcal{D}_n \mapsto \mathbb{R}^{\mathcal{Y}}$ which return decision rule $f_{\mathcal{D}_n}$.

Remark 1.1.10. Similarly, taking the $p-$norm ($p \geq 1$) of $R_{L,\nu}(f_{\mathcal{D}_n}) - R^*_{L,\nu}$ under measure $\mathbb{P}$, we can define the notions of strong upper rate (upper rate and lower minimax rate) of convergence in the space of $p-$integrable. Thus, a sequence of positive numbers $(a_n)_{n\geq 1}$ is called a strong upper rate in $L_p(\mathcal{Z}^n, \mathcal{B}(\mathcal{Z}^n), \mathbb{P})$ over prior class $\mathcal{P}$ if we have:

$$\limsup_{n\to\infty} \sup_{\mathbb{P}\in\mathcal{P}} \frac{\left\|R_{L,\nu}(f_{\mathcal{D}_n}) - R^*_{L,\nu}\right\|_p}{a_n} < \infty.$$

A sequence $(a_n)_{n\geq 1}$ is called upper convergence rate in $L_p\left(\mathcal{Z}^{\mathbb{N}}, \mathcal{B}\left(\mathcal{Z}^{\mathbb{N}}\right), \mathbb{P}\right)$ if

$$\lim_{n\to\infty} \sup_{\mathbb{P}\in\mathcal{P}} \frac{\left\|R_{L,\nu}(f_{\mathcal{D}_n}) - R^*_{L,\nu}\right\|_p}{a_n} < \infty.$$

Finally, a sequence $(a_n)_{n\geq 1}$ is called minimax (lower) rate of convergence in $L_p(\mathcal{Z}^n, \mathcal{B}(\mathcal{Z}^n), \mathbb{P})$ if

$$\liminf_{n\to\infty} \inf_{\mathcal{L}_n} \sup_{\mathbb{P}\in\mathcal{P}} \frac{\left\|R_{L,\nu}(f_{\mathcal{D}_n}) - R^*_{L,\nu}\right\|_p}{a_n} > 0.$$

Notice that if $(a_n)_{n\geq 1}$ is a (strong) upper rate over prior class $\mathcal{P}$ in expectation, then (by simply applying Markov's inequality) $(a_n)_{n\geq 1}$ is a (strong) upper rate over prior class $\mathcal{P}$ in probability. Furthermore, if $(a_n)_{n>1}$ is a lower minimax rate in probability then $(a_n)_{n\geq 1}$ is a lower minimax rate in expectation.

Generally, in the setting of statistical learning, we are interested in the algorithms which produce decision rules $f_{\mathcal{D}_n}$ whose upper convergence rates given by the Equation (1.3) match minimax lower convergence rates from Equation (1.4). The latter is assumed to hold either in probability of in $p-$norm. Furthermore, as has been done in Blanchard and Mücke (2018), in the learning rates for algorithms one

can track dependence on the additional parameters which specify the prior class, i.e. when $a_n := a_{n,\theta}$ where $\theta \in \Theta$ is some parameter set which specifies the set of available distributions $\mathcal{P}$. Furthermore, intuitively one would expect from the data-sample generated by non i.i.d. process with memory to contain "less" information and in this way lead to the worse rates than in the i.i.d. process scenario.

The first chapter of this thesis is devoted to the problem of statistical learning with least-square loss in the case where the measure $\mathbb{P}$ of random process $(Z_t = (X_t, Y_t))_{t\geq 1}$ satisfies certain weak-dependency assumption (the so-called $\tau-$mixing assumption see in Wintenberger (2010) also in Dedecker (1991)). We denote the risk of decision rule f under the squared loss by $R_{LS,\nu}(f) = \mathbb{E}_\nu\Big[(Y - f(X))^2\Big]$. When considering the problem of risk minimization as a stochastic optimization problem, one can readily check that a version of conditional expectation $f_\nu(x) = \int_{\mathcal{Y}} y\nu(y|x)$ (provided the distribution $\nu(\cdot|x)$ has finite second moment for μ almost all x) minimizes $R_{LS,\nu}(f)$ over all measurable functions $f : \mathcal{X} \mapsto \mathcal{Y}$. Furthermore, direct computation shows that, in this case, for any fixed measurable $f : \mathcal{X} \mapsto \mathbb{R}$ it holds that

$$R_{LS,\nu}(f) - R_{LS,\nu}(f_\nu) = \|f - f_\nu\|^2_{L_2(\mathcal{X},\mathcal{B}(\mathcal{X}),\mu)}.$$

Thus, in the case of least-squares loss the excess risk analysis of decision rule $f_{\mathcal{D}_n}$ reduces to the analysis of L_2 norm deviations of the difference $f_{\mathcal{D}_n} - f_\nu$.

Example 1.1.11. A standard model which illustrates the framework of statistical learning with least-squares loss is the regression problem. Namely, we consider the sequence $(X_t)_{t=1}^n$ to be a sequence of random variables with values in $\mathcal{X}$; assume $Y_t = f_*(X_t) + \varepsilon_t$, as the image of some unknown function $f_* : \mathcal{X} \mapsto \mathbb{R}$ corrupted by the noise sequence $(\varepsilon_t)_{t=1}^n$. We assume that $\mathbb{E}[\varepsilon_t] = 0$ and $\mathrm{Var}[\varepsilon_t|x] \leq \sigma^2 < \infty$ for $\mu-$ almost all x. In this model, the goal is to find a learning algorithm $\mathcal{L} : \mathcal{D}_n \mapsto \mathcal{Y}^{\mathcal{X}}$ which produces a measurable decision rule $\widehat{f}_{\mathcal{D}_n}$ which is a good estimate of f_*. The error of estimation is typically measured in $p-$norm ($p \geq 2$). It is easy to see that this model is a special case of the statistical learning scenario with squared loss when $f_* = f_\nu$, μ being the distribution of X_t and the conditional distribution $\nu(\cdot|x)$ being defined by the distribution of the noise $\varepsilon_t|x$. This setting is also sometimes referred to as *random design regression* (where the word "random" means that covariates x are generated from some distribution).

1.1.3 Introduction to learning with kernels

There are many works which are devoted to the problem of statistical learning with random observations and squared loss. In this short introduction, we focus mainly on the problem of estimating of the regression function f_ν by means of a decision rule in some reproducing kernel Hilbert space (RKHS). A Hilbert space of functions $\mathcal{H} := \{f : \mathcal{X} \mapsto \mathbb{R}\}$ equipped with an inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is called an RKHS if for every $x \in \mathcal{X}$ the evaluation functional $\delta_x(f) := f(x)$ is continuous in f. One can show that the continuity of the evaluation functional implies that the convergence in the norm of $\mathcal{H}$ implies point wise convergence of the elements $f \in \mathcal{H}$. Informally the latter means that if two objects from $\mathcal{H}$ are close (in the sense of norm $\|\cdot\|_{\mathcal{H}}$), then their values are close whatever the evaluation point is.

From the other end, we say that the function $k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ over domain $\mathcal{X}$ is a real-valued kernel if every matrix $K_n := (k(x_i, x_j))_{i,j=1}^n$ is positive semi-definite. It is known that the value of the kernel can be represented as an inner product in some Hilbert space H, namely $k\big(x, x^{'}\big) = \big\langle\phi(x), \phi(x')\big\rangle_H$. Some of the standard examples of kernels include (we consider here $\mathcal{X} \subset \mathbb{R}^d$): a **linear kernel** $k\Big(x, x^{'}\Big) = \Big\langle x, x^{'}\Big\rangle_{\mathbb{R}^d}$; a **polynomial kernel** of order $m \in \mathbb{N}$ $k_m\big(x, x^{'}\big) = \Big(\Big\langle x, x^{'}\Big\rangle_{\mathbb{R}^d} + c\Big)^m$; **Gaussian kernel** $k\big(x, x^{'}\big) = \exp\big(-\sigma^{-2}\big\|x - x^{'}\big\|_2^2\big)$, where $\sigma > 0$. In this case we call H a feature space and $\phi\big(\,\cdot\,\big)$ a feature map of kernel k. Lastly, we say that the RKHS $\mathcal{H}$ is generated by the kernel $k\big(\cdot,\cdot\big)$ if for every x it holds: $k_x := k(x,\cdot) \in \mathcal{H}$, and $f(x) = \big\langle f, k_x\big\rangle_{\mathcal{H}}$ for every $f \in \mathcal{H}$, and $x \in \mathcal{X}$ (i.e. the so-called reproducing property holds). In this case, we write $\mathcal{H}_k$ to denote the RKHS generated by

kernel $k(\cdot,\cdot)$, and say that the kernel $k(\cdot,\cdot)$ is a reproducing kernel of $\mathcal{H}_k$.

It is true (see Chapter 4 in Steinwart and Christmann (2008)) that for any kernel $k(\cdot,\cdot) : \mathcal{X}\times\mathcal{X}\mapsto\mathbb{R}$, there exists a unique RKHS $\mathcal{H}_k = \{f : \mathcal{X}\mapsto\mathbb{R}\}$ for which k is a reproducing kernel. The RKHS is in a certain sense, the smallest feature space H of the kernel k. Furthermore usage of kernel computationally is advantageous, since one does not need to know the feature map explicitly, but only the kernel product $k\big(x,x^{'}\big)$. From the other side, if one possesses any feature map $\phi(x) : \mathcal{X}\mapsto\mathcal{H}$ then, by the isometry property between an arbitrary feature space H and RKHS $\mathcal{H}_k$ for any $f\in\mathcal{H}_k$ one can write for $f(x) = \big\langle f, k_x\big\rangle_{\mathcal{H}_k} = \big\langle V^{-1}f, \phi(x)\big\rangle_H$ where $V : H\mapsto\mathcal{H}_k$ is metric surjection between H and $\mathcal{H}_k$ which always exists (see Theorem 4.21 in Steinwart and Christmann (2008)) and is defined as $Vg = \langle g, \phi(\cdot)\rangle_H$. In this case the problem of function evaluation is reformulated in terms of finding the inverse of V. The latter in general is computationally not easier than computing k_x and moreover the inverse is not necessarily unique; however, in some cases of kernels the map V can be defined explicitly and is an isometric isomorphism. For example (see Theorem 4.47 in Steinwart and Christmann (2008)) if $k_\sigma\big(x,x^{'}\big) = \exp\big(-\sigma^{-2}\big\|x-x^{'}\big\|_2^2\big)$ then we have $V_\sigma g = \big(\frac{2}{\pi}\big)^{\frac{d}{4}}\frac{1}{\sigma^{d/2}}\int_{\mathbb{R}^d}\exp\big(-\sigma^2\|x-\cdot\|_2^2\big)g(x)dx$ for $g\in L_2(\mathcal{X})$ (see Proposition 4.46 in Steinwart and Christmann (2008)).

In this thesis we restrict ourselves to the case of bounded and measurable kernels $k(\cdot,\cdot)$. Furthermore, for the statistical learning in the framework of non-parametric regression we assume that $f_\nu\in\mathcal{H}_k\subset L_2(\mathcal{X},\mathcal{B}(\mathcal{X}),\mu)$ which directly gives

$$\inf_{f\in\mathcal{H}} R_{LS,\nu}(f) = \inf_{f\in L_2(\mathcal{X},\mathcal{B}(\mathcal{X}),\mu)} R_{LS,\nu}(f) = R_{LS,\nu}(f_\nu).$$

Recall that in the learning setting we are generally interested in the excess risk $R_{LS,\nu}(f_{\mathcal{D}_n})-\inf_{f\in\mathcal{H}_k} R_{LS,\nu}(f)$ control, which under the assumption $f_\nu\in\mathcal{H}_k$ equals to $\|f-f_\nu\|_{L_2(\mathcal{X},\mu)}$. Notice that the latter assumption can be weakened by assuming that $\mathcal{H}_k$ is dense in $L_2(\mathcal{X},\mathcal{B}(\mathcal{X}),\mu)$ while $f_\nu\notin\mathcal{H}_k$ and preserving the last equality for the excess risk. The assumption $f_\nu\in\mathcal{H}_k$ is natural from the perspective of inverse problem theory as it is a necessary if one wants to study the behaviour of $f_{\mathcal{D}_n}-f_\nu$ in $\mathcal{H}_k-$ norm. In general the minimizer of $R_{LS,\nu}(f)$ over $\mathcal{H}$ does not necessarily exists. In this case define $f^+_{\mathcal{H}_k}$ to be the inclusion of the projection operator $P : L_2(\mathcal{X},\mathcal{B}(\mathcal{X}),\mu)\mapsto\overline{R}(I)$, where $\overline{R}$ is the closure of the range of the inclusion operator $I : \mathcal{H}_k\mapsto L_2(\mathcal{X},\mathcal{B}(\mathcal{X}),\mu)$ of the functions in $\mathcal{H}_k$ into $L_2(\mathcal{X},\mathcal{B}(\mathcal{X}),\mu)$. In the latter case we have that we can study $R_{LS,\nu}(f_{\mathcal{D}_n}) - R_{LS,\nu}\big(f^+_{\mathcal{H}_k}\big) = \big\|f_{\mathcal{D}_n}-f^+_{\mathcal{H}_k}\big\|^2_{L_2(\mathcal{X},\mu)}$.

One approach in non-parametric statistical learning which uses the machinery of RKHS and ensures existence of the solution of the problem of risk minimization is the usage of regularization. For example one can consider a regularized least-squares problem over the RKHS $\mathcal{H}_k$ which is generated by some measurable kernel k.

Example 1.1.12. Consider the following optimization criterion

$$\min_{f\in\mathcal{H}_k} R_{LS,\mathbb{P}_n}(f) + \lambda\|f\|^2_{\mathcal{H}_k} = \min_{f\in\mathcal{H}_k}\frac{1}{n}\sum_{t=1}^{n}(y_i - f(x_i))^2 + \lambda\|f\|^2_{\mathcal{H}_k}, \tag{1.5}$$

where $\lambda>0$. This is a known example of kernel ridge-regression problem (see Steinwart (2009),Caponetto and De Vito (2005) and Smola and Schölkopf (2002)). Let $f^\lambda_{\mathcal{D}_n}$ be a minimizer of Equation (1.5). Using the reproducing property and continuity of the evaluation functional, one can deduce the Representer Theorem (see Kimeldorf and Wahba (1970)). It states that the solution to every penalized optimization problem in Hilbert space has a unique minimizer which can be represented as a span of the kernel elements of the data sample. In particular, for the empirical penalized least-squares problem (1.5) one finds $f^\lambda_{\mathcal{D}_n} = \sum_{t=1}^n c_t k_{x_t}$, where $c\in\mathbb{R}^n$ is such that $c = (K_n + n\lambda\mathbb{I})^{-1}\mathbf{y}$, where $K_n = (k(x_i,x_j))_{i,j=1}^n$ is the kernel matrix and $\mathbf{y} = (y_1,\dots,y_n)$. By the law of large numbers for every fixed $f\in\mathcal{H}_k$, $R_{LS,\mathbb{P}_n}(f)$ converges to $R_{LS,\nu}(f)$. Therefore, one may consider empirical problem given in Equation (1.5) to be a

finite-sample estimate of the following population optimization problem

$$\min_{f\in\mathcal{H}_k} R_{LS,\nu}(f) + \lambda\|f\|^2_{\mathcal{H}_k}. \tag{1.6}$$

We refer to the Chapter 5 of the work Steinwart and Christmann (2008) for the questions of existence and uniqueness and of the dependence of the solutions to the regularized least-squares problems (1.5), (1.6) on the parameter $\lambda > 0$.

The effect of adding a regularization λ to the inverse of the kernel matrix in the least-squares regression problem can be seen as a particular way of filtering out the small eigenvalues of the matrix K_n and thus adding numerical stability to the solution $f^\lambda_{\mathcal{D}_n}$. This gives an intuition for a family of methods which act on the least-squares solution in a similar way (i.e. as a low-pass filter over the eigenvalues of the inverse of K_n). In the first chapter of this work, we study the properties of the generalized regularized learning schemes with values in the RKHS $\mathcal{H}_k$ which are based on the sample $\mathcal{D}_n$ which is generated by the so-called $\tau-$weakly-mixing process. To introduce them, consider a sample $\mathcal{D}_n = \{\mathbf{x}_t, y_t\}_{t=1}^n$, reproducing kernel $k(\cdot,\cdot)$ and the correspondent RKHS $\mathcal{H}_k$. Now define the following operators:

$$S_n : \mathcal{H}_k \mapsto \mathbb{R}^n, \quad (S_n h)_j = \langle h, k_{\mathbf{x}_j}\rangle = h(\mathbf{x}_j), \qquad S^*_{\mathbf{x}} : \mathbb{R}^n \mapsto \mathcal{H}_k, \quad S^*_n\mathbf{y} = \frac{1}{n}\sum_{j=1}^n y_j k_{\mathbf{x}_j},$$

$$T_n := S^*_n S_n : \mathcal{H}_k \mapsto \mathcal{H}_k, \quad T_n h = \frac{1}{n}\sum_{j=1}^n k_{\mathbf{x}_j}\langle k_{\mathbf{x}_j}, h\rangle, \qquad L_n := S_n S^*_n : \mathbb{R}^n \mapsto \mathbb{R}^n, \quad L_n = n^{-1}K_n,$$

where K_n is the kernel matrix as above. Lastly, let S, S^*, T, L be the corresponding population analogs of the empirical operators, namely:

$$\begin{aligned}
&S : \mathcal{H}_k \mapsto L_2(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mu), && Sh = [f]_\mu, \text{ such that } \langle h, k_x\rangle = [f]_\mu(x), \mu - \text{ a.s}\\
&T : \mathcal{H}_k \mapsto \mathcal{H}_k, && Th = \int_{\mathcal{X}} k_z\langle k_z, h\rangle\mu(dz),\\
&L : L_2(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mu) \mapsto L_2(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mu), && Lf(x) = \int_z k(x,z)f(z)\mu(dz), \mu \text{ a.s.}
\end{aligned}$$

Notice that with the above notations, the solutions of both the empirical and population versions of the least-squares penalized problem given by Equations(1.5) and (1.6) respectively can be written as $f^\lambda_{\mathcal{D}_n} = (T_n + \lambda\mathbb{I})^{-1}S^*_n\mathbf{y}$ and $f_\lambda = (L + \lambda\mathbb{I})^{-1}Lf_\nu$. In Chapter 2 we consider the following decision rules, which are generated by a family of operator-valued maps $g_\lambda : \mathcal{H}_k \mapsto \mathcal{H}_k$ (the so-called *regularization*):

$$f^\lambda_{\mathcal{D}_n} = g_\lambda(T_n)S^*_n\mathbf{y}, \tag{1.7}$$

where $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$. This estimator is also sometimes referred to as the bandpass filter as its action can be seen as a transformation of the eigenvalues of the covariance operator T_n. For example, the choice $g_\lambda(t) = \frac{1}{t+\lambda}$ corresponds to the solution of the least-squares penalized regularization problem as given by Equation (1.5) (also called Tikhonov regularization or ridge regression, see for example in Caponetto and De Vito (2005), and Fischer and Steinwart (2017)); $g_\lambda(t) = \sum_{i=0}^{\lambda^{-1}}(1-t)^j$ to Landweber iteration (also called gradient descent, see ex. Robbins (1952), Pillaud-Vivien et al. (2018)); and $g_\lambda(t) = t^{-1}\mathbb{I}_{t\geq\lambda}$ corresponds to the spectral cut-off (see Rosasco et al. (2010) and Blanchard et al. (2007)).

Motivation for the choice of regularization comes from the theory of inverse problems (with fixed observations) in which, under certain conditions on the regularization function (see Bauer et al. (2009), Engl et al. (2000)), the population version of Equation (1.7) (i.e. the function which is obtained through the im-

age of the integral and adjoint to the evaluation operator with respect to measure $\mathbb{P}$ over $\left((\mathcal{X}\times\mathcal{Y})^{\mathbb{N}},\mathcal{B}\left((\mathcal{X}\times\mathcal{Y})^{\mathbb{N}}\right)\right)$ converges to the minimizer of the population optimization problem.

Notice that although in the framework of least-squares regression there are universally consistent rules for the problem of least-squares regression but, because of the existence of the variant of the No-free-lunch Theorem (see Corollary 6.8 in Steinwart and Christmann (2008) or Chapter 3 in Győrfi (2002)), it is impossible to obtain learning rates without additional assumptions on the marginal distribution ν even in the case when $\mathbb{P}$ is a product measure. In the regression case, one of the possible assumptions on ν translates into a notion of the complexity of the functional class to which the regression function f_ν belongs. One way to obtain such complexity bounds is to introduce smoothness assumptions on the regression function f_ν (and thus on the underlying functional class). In Chapter 2 we pose such assumption on f_ν in terms of the so-called standard Hölder source condition for the linear embedding problem. From the perspective of statistical learning with the least-squares loss, the main contribution of Chapter 2 is establishing strong upper rates of convergence of the general Hilbert-valued learning algorithms given by Equation 1.7 when training sample $\mathcal{D}_n$ comes from a trajectory of some process with decaying correlation assumptions on the distribution. Depending on the correlations strength we highlight the scenario in which upper convergences rates are essentially optimal (meaning that up to a logarithmic term in the number of observations they match the minimax rate convergence rates in the i.i.d. scenario). We highlight the cases when convergence rates are worse by polynomial factor and point out that this only the consequence of suboptimal rates in the correspondent deviation inequalities. We give more details on the framework of statistical learning with dependent random observations in the next section.

1.2 Statistical learning from dependent random observations.

1.2.1 Introduction to the notion of asymptotic independence (weak-dependence).

To be able to navigate somewhat consciously between the notions of dependence, I introduce a general framework of dependence measures from somewhat functional perspective. The spirit of this is taken from the introduction of the book Dedecker and Merlevede (2015) and from the survey of Bradley (2002), however, to the best on my knowledge, general scope of this presentation seems to be new. We use the notation $\mathbb{E}_{\mathcal{A}}[\cdot]:=\mathbb{E}[\cdot|\mathcal{A}]$ to denote conditional expectation under regular conditional probability $\mathbb{P}(\cdot|\mathcal{A})$ for any $\sigma-$algebra $\mathcal{A}\subset\mathcal{F}$. Recall that two $\sigma-$algebras $\mathcal{A},\mathcal{B}\subset\mathcal{F}$ are said to be independent if for all $A\in\mathcal{A}$, $B\in\mathcal{B}$ it holds that $\mathbb{P}(A\cap B)=\mathbb{P}(A)\mathbb{P}(B)$. Alternatively, the last expression can be written as

$$\sup_{A\in\mathcal{A},B\in\mathcal{B}}\mathrm{Cov}_{\mathbb{P}}[\mathbb{I}_A,\mathbb{I}_B]=0, \tag{1.8}$$

where $\mathbb{I}_A(\omega)=\mathbb{I}_{\omega\in A}$ is the standard notation for the indicator random variable. Extending (1.8) by linearity of the covariance form to the set of simple functions, and using the fact that every bounded r.v. is a pointwise (in ω) limit of elements of the set of simple random variables, we can write that $\mathcal{A}$, $\mathcal{B}$ are independent iff

$$\sup_{f\in L_\infty(\Omega,\mathcal{A},\mathbb{P}),h\in L_\infty(\Omega,\mathcal{B},\mathbb{P})}\mathrm{Cov}_{\mathbb{P}}[f,h]=0.$$

As we will demonstrate below, this form of covariance condition will be the most convenient in many cases of usage to control dependence between "past" and "future" of stochastic process which is represented by the information contained in $\mathcal{A}$ and $\mathcal{B}$ correspondingly. Informally, it is natural to say that $\sigma-$algebras $\mathcal{A},\mathcal{B}$ are "almost independent" if

$$\sup_{f\in\mathcal{G},h\in\mathcal{H}}|\mathrm{Cov}[f,h]|<\delta$$

for $\delta \to 0$ and some apriory chosen classes $\mathcal{G} \subset L_\infty(\Omega, \mathcal{A}, \mathbb{P})$ and $\mathcal{H} \subset L_\infty(\Omega, \mathcal{B}, \mathbb{P})$. For any classes $\mathcal{G}, \mathcal{H}$ of $(\mathcal{A}, \mathcal{B}(\mathbb{R}))$ and $(\mathcal{B}, \mathcal{B}(\mathbb{R}))$ measurable bounded real maps correspondingly, $\mathcal{G} = \{f : \Omega \mapsto \mathbb{R}\}$, $\mathcal{H} = \{h : \Omega \mapsto \mathbb{R}\}$ such that point-wise limits of any sequence belong to $L_\infty(\Omega, \mathcal{A}, \mathbb{P})$, $L_\infty(\Omega, \mathcal{B}, \mathbb{P})$, the latter condition is equivalent to $\sup_{f \in L_\infty(\Omega, \mathcal{A}, \mathbb{P}), h \in L_\infty(\Omega, \mathcal{B}, \mathbb{P})} |\mathrm{Cov}[f, h]| < \delta$. Varying the functional classes $\mathcal{G}, \mathcal{H}$ of $\mathcal{A}$, $\mathcal{B}$ measurable functions correspondingly (non necessarily being the subsets of essentially bounded functions), under the additional assumptions that f, h, fh are integrable (which, for example, is fulfilled if $f \in L_p(\Omega, \mathcal{A}, \mathbb{P})$,$h \in L_q(\Omega, \mathcal{B}, \mathbb{P})$ with $p^{-1} + q^{-1} \leq 1$) we can define the following dependence measure

$$a(\mathcal{G}, \mathcal{H}, \mathbb{P}) := \sup_{(f,h) \in \mathcal{G} \times \mathcal{H}} |\mathrm{Cov}_{\mathbb{P}}[f, h]|, \tag{1.9}$$

where we take the supremum over f, g such that $f : (\Omega, \mathcal{A}) \mapsto (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $h : (\Omega, \mathcal{B}) \mapsto (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. This notion of dependence measure naturally translates to the case of two random variables. Namely, if X, Y are random variables over $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ and $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ and their joint distribution $\nu_{X,Y}$ over $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}(\mathcal{X} \times \mathcal{Y}))$ then we define the dependence measure between random variables X,Y as the dependence measure between corresponding sigma-fields $\mathcal{B}(\mathcal{X})$ and $\mathcal{B}(\mathcal{Y})$ under measure $v_{X,Y}$.

Example 1.2.1 (**Mixing coefficients**). In what follows we introduce the notion of mixing coefficients. We deviate a bit from the common way of presenting mixing coefficients (see Bradley (2002), Dedecker and Prieur (2005)) as a supremum over the difference between joint probability law and the product of its marginal over the certain classes of events and propose (in some sense) a more functional-analytic approach. This corresponds in spirit to the particular cases of dependence measure $a(\cdot, \cdot, \mathbb{P})$ introduced in the previous paragraph. We firstly define the following dependence measures between $\sigma-$fields as follows:

$$\begin{aligned}
\alpha(\mathcal{A}, \mathcal{B}) &= \sup_{A \in \mathcal{A}, B \in \mathcal{B}} |\mathrm{Cov}[\mathbb{I}_A, \mathbb{I}_B]| = \sup_{\substack{f \in L_\infty(\mathcal{A}), h \in L_\infty(\mathcal{B}) \\ \|f\|_\infty \leq 1 \|h\|_\infty \leq 1}} |\mathrm{Cov}[f, h]| \\
\rho(\mathcal{A}, \mathcal{B}) &= \sup_{\substack{f \in L_2(\mathcal{A}), h \in L_2(\mathcal{B}) \\ \|f\|_2 \leq 1, \|h\|_2 \leq 1}} |\mathrm{Cov}[f, h]| \\
\phi(\mathcal{A}, \mathcal{B}) &= \sup_{A \in \mathcal{A}, B \in \mathcal{B}, \mathbb{P}(A) > 0} |\mathbb{E}[\mathbb{I}_B | A] - \mathbb{E}[\mathbb{I}_B]| = \sup_{A \in \mathcal{A}, B \in \mathcal{B}, \mathbb{P}(A) > 0} |\mathrm{Cov}[\mathbb{I}_A / \|\mathbb{I}_A\|_1, \mathbb{I}_B]|
\end{aligned} \tag{1.10}$$

We define the $\alpha, \rho, \phi-$mixing coefficient of the process $(X_t)_{t \in \mathbb{Z}}$ over $(\Omega, \mathcal{A}, \mathbb{P})$ as

$$\alpha(k) = \sup_{i \in \mathbb{Z}} \alpha\big(\mathcal{F}^i_{-\infty}, \mathcal{F}^\infty_{i+k}\big),$$

$$\rho(k) = \sup_{i \in \mathbb{Z}} \rho\big(\mathcal{F}^i_{-\infty}, \mathcal{F}^\infty_{i+k}\big),$$

$$\phi(k) = \sup_{i \in \mathbb{Z}} \phi\big(\mathcal{F}^i_{-\infty}, \mathcal{F}^\infty_{i+k}\big)$$

where we define $\mathcal{F}^b_a = \sigma(X_t : a \leq t \leq b)$.

Notice that dependence measure of Equation 1.10 can be presented as the bi-linear covariance form by means of coefficient a as in Equation (1.9). Namely, for σ fields $\mathcal{A}, \mathcal{B} \subset \mathcal{F}$, $1 \leq p, q \leq \infty$ and $L_p(\Omega, \mathcal{A}, \mathbb{P})$, $L_q(\Omega, \mathcal{B}, \mathbb{P})$, being the standard equivalence classes of $p-$integrable functions with respect

to measure $\mathbb{P}$, we can readily check that the following holds

$$\alpha(k) = \sup_i a\big(B_{L_\infty\left(\Omega,\mathcal{F}^i_{-\infty},\mathbb{P}\right)}(1), B_{L_\infty\left(\Omega,\mathcal{F}^\infty_{i+k},\mathbb{P}\right)}(1), \mathbb{P}\big)$$
$$\rho(k) = \sup_i a\big(B_{L_2\left(\Omega,\mathcal{F}^i_{-\infty},\mathbb{P}\right)}(1), B_{L_2\left(\Omega,\mathcal{F}^\infty_{i+k},\mathbb{P}\right)}(1), \mathbb{P}\big)$$
$$\phi(k) = \sup_i a\big(B_{L_1\left(\Omega,\mathcal{F}^i_{-\infty},\mathbb{P}\right)}(1), B_{L_\infty\left(\Omega,\mathcal{F}^\infty_{i+k},\mathbb{P}\right)}(1), \mathbb{P}\big)$$

The theory of mixing processes dates back to the works Rosenblatt (1956), Ibragimov (1959) where the concepts of α and $\phi-$ mixing coefficients were introduced. $\alpha-$mixing coefficients are also sometimes referred to as strong mixing coefficients. The main motivation, which gave rise to its development, were the needs of statistical inference for the processes which do not have a specific functional structure (like AR-models or Gaussian processes for instance), but possess a certain kind of decaying correlation property between past and future which in the limit case (i.e. when the time gap between past and future goes to infinity) recovers processes with independent noise component processes.

The concept of the $\rho-$mixing coefficient was introduced in Kolmogorov and Rozanov (1960). Furthermore, notions of $\alpha, \rho, \phi-$ mixing were widely studied by Bradley et al. (1987), Dehling and Philipp (1982), Peligrad (1983),Bradley (2007). Apart from these three examples, in the theory of mixing processes there are many other concepts of dependence which characterize the decaying correlations between whole past and whole future of the process, for example $\beta-$mixing (McDonald et al. (2015)) , $\eta-$mixing (Kontorovich (2006)), $\psi-$mixing (Bradley (2007)).

We refer to the survey paper of Bradley (2002) for further notions of mixing, their history and further properties and relations to the other dependence measures.

Remark 1.2.2. The advantage of this presentation is that for $\mathrm{Cov}[f, g]$, one can directly apply techniques from functional analysis to compare these measures of dependence and derive covariance inequalities. Furthermore, we notice that if we, say, fix parameter q, then the smaller we choose p^{-1} , the broader the functional class (and thus a weaker notion of the mixing coefficient) we obtain.

From the work Andrews (1984) (see p.9 therein) it is known that there exists a stochastic process with very simply structure which is not $\alpha-$mixing. Namely, one can consider a simple chain $X_n = \frac{1}{2}(X_{n-1} + \varepsilon)$, X_0 is independent of $(\varepsilon_i)_{i\geq 1}$ and $\varepsilon_i = \mathcal{B}\big(\frac{1}{2}\big)$. One can show that $\alpha(\sigma(X_0), \sigma(X_n)) = \frac{1}{2}$, thus $(X_n)_{n\geq 1}$ is not even $\alpha-$mixing. In this case, a natural idea is to relax the mixing assumption and to consider the weaker classes of processes obtained by functional transforms.

Example 1.2.3. Functional weak dependence Let $\mathcal{C}$ be a Banach space of real bounded functions $\{f : \mathcal{X} \mapsto \mathbb{R}\}$ equipped with the norm $\|\cdot\|_{\mathcal{C}} = \|\cdot\| + C(\cdot)$, where $C(\cdot)$ is some seminorm over $\mathcal{X}$ and $\|\cdot\|$ is a standard supremum norm on $\mathcal{C}$. Denote $\mathcal{C}_1 := \{f \in \mathcal{C}, C(f) \leq 1\}$. Now, for a stochastic process $(X_t)_{t\in\mathbb{N}}$, arbitrary $i \in \mathbb{Z}$ and fixed $k \in \mathbb{N}$, consider $\mathcal{A}_i = \sigma(X_j : j \leq i)$, $\mathcal{B}_{i+k} = \sigma(X_{i+k})$, $\mathcal{G} = L_1(\Omega, \mathcal{A}_i, \mathbb{P})$, $\mathcal{H} = \mathcal{C}_1 \circ L_\infty(\Omega, \mathcal{B}_{i+k}, \mathbb{P})$ and define the following coefficient :

$$\phi_{\mathcal{C}}(k) := \sup_i a(\mathcal{G}, \mathcal{H}, \mathbb{P})$$

which can be written explicitly written (using X to denote any $(\mathcal{A}_i, \mathcal{B}(\mathcal{X}))$ measurable integrable random variable and $Y := X_{i+k}$).

$$\phi_{\mathcal{C}}(k) := \sup_{i\in\mathbb{N}}\{|\mathrm{Cov}[X, f(Y)]|, \|X\|_{L_1(\Omega,\mathcal{A}_i,\mathbb{P})} \leq 1, f \in \mathcal{C}_1\}.$$

The latter condition is introduced in Maume-Deschamps (2006) and extends the ideas of weak-dependent coefficients by Dedecker and Prieur (2005), Doukhan and Louhichi (1999). We consider the case of stochastic process valued in some bounded set $\mathcal{X} \subset \mathbb{R}$ and consider the Lipschitz and total variation

seminorms:

$$\Lambda_M := B_{Lip}(M)\left\{ f : \mathcal{X} \mapsto \mathbb{R}, \sup_{x,y\in\mathcal{X}} \frac{|f(x)-f(y)|}{|x-y|} \leq M \right\}$$

$$BV_M := B_{BV}(M)\left\{ f : \mathcal{X} \mapsto \mathbb{R}, \|f\|_{TV} := \sup_{x_0,\dots,x_n\in\Delta\subset\mathcal{X}} \sum_{i=1}^{n} |f(x_i)-f(x_{i-1})| \leq M \right\}$$

Notice that the notion of Lipschitz semi-norm and Lipschitz functional class can be trivially extended to the case of general normed spaces. Weakly $\tau(\cdot), \phi(\cdot)$ coefficients (see Wintenberger (2010), Dedecker and Merlevede (2015)) are obtained when considering $\mathcal{C}_1 = \Lambda_1$, while if we take $\mathcal{C}_1 = BV_1$ we obtain the so-called weak $\phi-$mixing coefficients (see Rio (1996), Dedecker and Merlevede (2015)).

1.2.2 Projective dependence measure. Mixingales

A stronger approach is to have the control over the conditional expectation of a stochastic process instead of having control of the covariance. The latter leads to the projective type dependence criterion. Let X, Y be centered real random variables over the same space $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$, $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}$ such that X is $(\mathcal{A}, \mathcal{B}(\mathcal{X}))$-measurable while Y is $(\mathcal{B}, \mathcal{B}(\mathcal{Y}))$ -measurable. We have

$$\mathrm{Cov}[X,Y] = \mathbb{E}[X\mathbb{E}_{\mathcal{A}}[Y]] = \mathbb{E}[Y\mathbb{E}_{\mathcal{B}}[X]]$$

from which, by applying Hölder's inequality, we deduce that

$$|\mathrm{Cov}[X,Y]| \leq \|X\|_{L_p(\Omega,\mathcal{A},\mathbb{P})} \|\mathbb{E}_{\mathcal{A}}[Y]\|_{L_q(\Omega,\mathcal{B},\mathbb{P})} \qquad |\mathrm{Cov}[X,Y]| \leq \|Y\|_{L_q(\Omega,\mathcal{B},\mathbb{P})} \|\mathbb{E}_{\mathcal{B}}[X]\|_{L_p(\Omega,\mathcal{A},\mathbb{P})}.$$

These inequalities give rise to the following definition of projective weak-dependence measure.

Definition 1.2.4. For any probability space $(\Omega, \mathcal{F}, \mathbb{P})$, $\mathcal{A}, \mathcal{B} \subset \mathcal{F}$, some value $1 \leq q \leq \infty$, and some class F of centered $(\mathcal{B}, \mathcal{B}(\mathbb{R}))$-measurable and $q-$ integrable random variables we define:

$$\mathcal{E}_q(\mathcal{A}, F) = \sup_{X\in F} \|\mathbb{E}_{\mathcal{A}}[X]\|_q. \tag{1.11}$$

For the case of stochastic process $(X_t)_{t\in\mathbb{N}}$ we define the following "projective" dependence coefficients by varying the $\sigma-$algebras $\mathcal{A}, \mathcal{B}$ and classes F. Namely, for $i, k \in \mathbb{Z}$ we set $\mathcal{B} = \mathcal{F}_{i+k}^{\infty}$ and define $e_{q,L_q}(k) := \sup_{i\in\mathbb{Z}} \mathcal{E}_q\big(\mathcal{F}_{-\infty}^{i}, L_q(\Omega, \mathcal{B}, \mathbb{P})\big)$.

Remark 1.2.5. The intuition behind the notion "projective" dependence is apparent from the definition, as one considers the variation of the $q-$norm of the orthogonal projection of the variable X on the subspace generated by $\mathcal{A}-$measurable functions. Furthermore, from the application of the Hölder's inequality above and by considering the notion of dependence coefficient $a(\cdot,\cdot,\mathbb{P})$, it follows that for every $1 \leq p, q \leq \infty$

$$a\big(B_{L_p(\Omega,\mathcal{A},\mathbb{P})}(1), B_{L_q(\Omega,\mathcal{B},\mathbb{P})}(1), \mathbb{P}\big) \leq e_q(k)$$

for every $k \in \mathbb{N}$.

The following Lemma (see Lemma 1.1.2 in Maume-Deschamps (2006)) allows one to relate a projective-type dependence measure (with certain classes F and norm $q = \infty$) to the notion of functional dependence, and it will be important for further analysis in Chapter 2.

Lemma 1.2.6. *For the $\mathcal{X}-$valued stationary stochastic process $(X_t)_{t\in\mathbb{N}}$, the functional class of real-valued and bounded functions $\mathcal{C} = \{f : \mathcal{X} \mapsto \mathbb{R}\}$ as in Example 1.2.3 such that $\mathbb{E}[f(X_t)] = 0$ for any*

$f \in \mathcal{C}_1$. *Consider* $\sigma-$*field* $\mathcal{F}_b^a$, *a weak-dependency coefficient* $\phi_{\mathcal{C}}(k)$ *all being as defined above. It holds:*

$$\phi_{\mathcal{C}}(k) = \sup_{i\in\mathbb{Z}} \sup_{g\in\mathcal{C}_1} \left\| \mathbb{E}\left[g(X_{i+k})|\mathcal{F}_{-\infty}^i\right] \right\|_\infty. \tag{1.12}$$

Remark 1.2.7. One can easily see that if in the Definition 1.12 we take $q = \infty$, $\mathcal{A} = \mathcal{F}_{-\infty}^i$, $\mathcal{B} = \sigma(X_{i+k})$, $F = \mathcal{C}_1 \circ L_1(\Omega, \mathcal{F}, \mathbb{P})$ then by Lemma 1.2.6 we have $\phi_{\mathcal{C}}(k) = e_{\infty,\mathcal{C}\circ L_1}(k)$. The latter shows that certain projective-type dependence measures (and thus functional weak-dependence measures) induce the notion of functional weakly-dependent mixing coefficients. Recall notation Λ_1, BV_1 for the unit balls in the spaces of Lipschitz and functions with bounded variation. For a real-valued stationary stochastic process $(X_t)_{t\geq 1}$,$\|X_t\|_\infty \leq 1$ consider $L_1^0(\mathbb{P}_X)$ to be a class of functions such that $\mathbb{E}[f(X)] = 0$, $\sigma-$field $\mathcal{F}_{-\infty}^i$ as before, we denote $P_X = \mathbb{P} \circ X_t$ (see Dedecker et al. (2007) Dedecker and Merlevede (2015) for these and another examples) and obtain

$$\begin{aligned}
\tilde{\alpha}(k) &= \sup_i \sup_{f\in BV_1\cap L_1^0(\mathbb{P}_X)} \left\| \mathbb{E}\left[f(X_{i+k})|\mathcal{F}_{-\infty}^i\right] \right\|_1 = e_{1,BV_1}(k) \\
\tilde{\phi}(k) &= \sup_{i\in\mathbb{N}} \sup_{f\in BV_1\cap L_1^0(\mathbb{P}_X)} \left\| \mathbb{E}\left[f(X_{i+k})|\mathcal{F}_{-\infty}^i\right] \right\|_\infty = e_{\infty,BV_1}(k) \\
\tilde{\tau}(k) &= \sup_{i\in\mathbb{N}} \sup_{f\in\Lambda_1\cap L_1^0(\mathbb{P}_X)} \left\| \mathbb{E}\left[f(X_{i+k})|\mathcal{F}_{-\infty}^i\right] \right\|_\infty = e_{\infty,\Lambda_1}(k).
\end{aligned}$$

In this thesis, we mainly work with two types of weak-dependency measures: the functional weak-dependency coefficient $\phi_{\mathcal{C}}(\cdot)$ in Chapters 2 and 4 and the extension of the projective-dependence measure $e_p(\cdot)$ to the case of random fields (see Chapter 5), which can also be seen as an extension of the concept of mixingale whose short definition is presented below.

Mixingale In all previous examples, to characterize the dependence we always considered the canonical filtration generated by the stochastic process $(X_t)_{t\mathbb{N}}$. Following Mc Leish (1975) and Andrews (1988) (see also the Definition 1.1 in Dedecker and Merlevede (2015)), consider some increasing sequence of $\sigma-$algebras such that for any n, we have $\mathcal{F}_n \subset \mathcal{F}$ and $p \geq 1$. The sequence $(X_t, \mathcal{F}_t)_{t\geq 1}$ is called an L_p-mixingale if there exist non-negative sequences $(c_n)_{n\in\mathbb{Z}}$ and $(\psi_n)_{n\in\mathbb{Z}}$ such that $\psi_n \to 0$ as $n \to \infty$, and for all $n \in \mathbb{Z}$ we have

$$\begin{aligned}
\|X_n - \mathbb{E}[X_n|\mathcal{F}_{n+k}]\|_{L_p(\Omega)} &\leq c_n\psi_{k+1}, \\
\|\mathbb{E}[X_n|\mathcal{F}_{n-k}]\|_{L_p(\Omega)} &\leq c_n\psi_k.
\end{aligned} \tag{1.13}$$

Notice that when $\mathcal{F}_n = \sigma(X_i : i \leq n)$, mixingales satisfies the projective-type criterion of Example 1.2.2. Namely, in this case the first condition in (1.13) is straightforwardly satisfied (as X_n is $\mathcal{F}_{n+k}-$measurable), and the second condition becomes a form of a projective-type dependency condition. Under this general assumption, many processes (which are non-mixing) can be classified. However, obtaining general moment inequalities or limit theorems are typically difficult in this scenario (see (Peligrad et al., 2006, 2007) for some results for general stochastic processes which have a martingale-like component; see also Doukhan (1994) for more examples of the processes which are mixingale like).

1.2.3 Concentration inequalities for weakly-dependent processes.

For a random variable $Z : \Omega \mapsto \mathbb{R}$, the question of establishing inequalities of type $\mathbb{P}(Z \geq \mathbb{E}[Z] + \varepsilon)$ (i.e. concentration inequalities) is a fairly general and vast topic of interest both in probability and statistics. Such a phenomenon is vastly studied in the context of product measure $\mathbb{P}$ on the product space $(\Omega, \mathcal{F})$ where $\Omega = \left(\Omega^{'}\right)^n, \mathcal{F} = \mathcal{B}\left(\left(\Omega^{'}\right)^n\right)$ (see Talagrand (1995), Ledoux (2001), McDiarmid (1989)) under quite general assumptions on the "smoothness" properties of the underlying function Z. Most of

the techniques use either information-theoretic (see ex. Marton (2004)), isoperemetric ((Ledeoux, 1997; Ledoux, 2001)), or martingale-difference approaches (see Pinelis (1994)), or exploits in a different way the structure of the underlying variable Z. The situation becomes more complex if the measure $\mathbb{P}$ is a non-product measure, as in this case one needs to quantify the dependency between the marginals of $\mathbb{P}$. Several works has been done to classify the dependence between the marginals of the distribution of the stochastic process. These, as it was discussed in the previous section include the notion of mixing coefficients, the weak-dependency assumption (see (Dedecker, 1991) andvDedecker et al. (2007)) or the functional weak-dependency assumption Maume-Deschamps (2006). The problem of establishing even the asymptotic results for non-product measures is already interesting in the case when $Z = \sum_{t\in\mathcal{T}} X_t$, $\mathcal{T}$ is some subset of some vector space and $(X_t)_{t\in\mathcal{T}}$ is a stochastic process with values Banach space $(\mathcal{B}, \|\cdot\|)$. Standard Hoeffding's and Bernstein's inequalities for general norms of sums of i.i.d. random vectors when $\mathcal{T} \subset \mathbb{N}$ due to Pinelis and Sakhanenko (1986) (which recovers the well-known Hoeffding's and Bernstein's inequalities in the case of bounded real-valued random variables) are extended to the case of concentration of super-martingales in the work Pinelis (1992) and more general functions with domain in some Banach space in Pinelis (1994).

Beyond the (super)martingale setting, the need to handle more general processes which have some "asymptotic independence" assumptions led to the concept of mixing, weakly-dependent processes and mixingales whose exact definitions were given in the previous section in terms of dependence coefficients $a(\cdot,\cdot,\mathbb{P})$ or $e(\cdot)$. In these settings, mostly martingale-like and coupling techniques were used and improved and combined with other methods to obtain concentration inequalities for the sums of *real-valued* dependent random variables. Most of the techniques relies on the splitting of the data-sample into blocks (using some practical way) and considering samples from different blocks as "almost" independent (up to a contamination term which arises when substituting the joint probability distribution by product of marginals). In such spirit generalizations of Bernstein's inequality for ϕ-mixing random processes were obtained in Samson (2000); also Bernstein-type inequalities for geometrically $\alpha-$mixing processes and moderate deviation principles were derived in Merlevede et al. (2009); deviation inequalities for real-valued sums of variables from general $\alpha-$mixing processes were obtained in Bosq (1993). In Kontorovich and Ramanan (2008), the martingale difference method is used to establish general McDiarmid-type concentration inequalities for real-valued Lipschitz functions of dependent random sequences on a countable state space. Using logarithmic Sobolev inequalities and the contractivity condition related to Dobrushin and Shlosman's strong mixing assumptions, general non-product measure concentration inequalities were obtained in Marton (2004). We relate to a proper discussion on the subject of existing results on the concentration of weakly-dependent partial sums of random process to Chapter 2.

The goal of the first part of the thesis is to consider the general class of weakly $\mathcal{C}-$mixing processes and derive new concentration inequalities of Bernstein-type for the deviations of the random sums of Banach-valued random variables. In the Chapter 5 we extend the projective type-criteria (formulated in terms of dependency measure $\mathcal{E}_q(\mathcal{A}, F)$) to the case of weakly dependent random fields and derive deviation bounds (on the exponential scale in probability and of Burkholder type in $p-$norm) for the partial sums.

1.2.4 Statistical learning with dependent random observations

This part is devoted to a discussion of an extension of the concept of statistical learning from examples to the case of non i.i.d. data observations. Statistical motivation for investigation of this framework can be described in various perspectives. Firstly, stochastic independence of random observations in many cases is an assumption rather than a property which can be inhereted from some empirical experiment and often hard to verify for data sources with complex structure. Thus in the framework of learning from observation one wants to extend the dependence notion in order to include processes which which samples can be seen as "nearly independent" Secondly, in such an extension one wants to consider different dependence measures, that is to give mathematical classification for the range of processes

for which the samples are "strongly" or "weakly" dependent. Lastly, having a certain class of such dependent processes, in the framework of learning from examples one wants to investigate how the desirable stochastic independence properties (in terms of rates for either excess risk or generalization error) are preserved or in which way they are contaminated (due to the presence of dependencies) in the learning rates of the algorithms when using dependent data sample for training.

The setting of statistical learning with weakly-dependent or mixing observations is definitely not new. In the work by Hang and Steinwart (2017), the authors consider the general notion of $\mathcal{C}-$mixing processes as given in Equation (1.12). They also develop probabilistic toolbox (in terms of an exponential Bernstein-type inequality for sums of real-valued random variables) and use it to establish learning rates for the LS-SVM algorithm (see Steinwart and Christmann (2008)) in the setting of nonparametric least-squares regression. In particular, when the regression function belongs to the bounded ball in the space $B^t_{2s,\infty}(\mathcal{X})$ (Besov space of smoothness t, $s \geq 1$), LS-SVM leads to the essentially (up to a small polynomial factor) optimal rates when $t > \frac{d}{2}$.

When finding the upper bounds for the excess-risk of the data-dependent decision rule $f_{\mathcal{D}}$, one aims to control the generalization error, i.e. the difference $R_{L,P}(f) - R_{L,P_n}(f)$ uniformly over $f \in \mathcal{F}$. Such control leads to the notion of class complexity (VC dimension, Rademacher complexity or covering number). For example, the generalization of the bounds for Rademacher complexity to the case of non i.i.d. random samples (more precisely under the so-called $\beta-$mixing assumption which implies $\alpha-$mixing) is provided in Mohri and Rostamizadeh (2008)); the framework of structured risk minimization and VC dimension is adapted to the ergodic time series and uniform convergence results (i.e. laws of large numbers which hold uniformly over the functional class of prediction rules $\mathcal{F}$) are obtained for the processes which fulfill certain algebraic or exponential mixing condition by Meir (2000). In the work of Vidyasagar (2003), author discusses the sufficient conditions on the underlying data-generating process which ensures that the uniform convergence results can be transferred from the setting with i.i.d. random observations to the processes with dependencies (and thus become uniform ergodic theorems). In the work Adams and Fournier (2003), the uniform law of large numbers is shown for the ergodic processes with values in complete separable metric spaces and over functional classes with finite VC dimension.

From the stability perspective, learning rates for various classes of learning algorithms are analysed for the case $\beta-$ and $\varphi-$ mixing sequences by Mohri and Rostamizadeh (2010). In this framework one considers the generalization bounds for some algorithm $\mathcal{A}$ which depend on the sensitivity to the changes of the expected risk if one point is being changed in the training sample. This work shows that a "stable" algorithm provides good generalization results in the case when $\beta-$mixing coefficient decays sufficiently fast. All of the results therein assume that the underlying data generating process satisfies certain mixing condition and demands the knowledge of the mixing rate (which is not given in practice). In McDonald et al. (2011), McDonald et al. (2015), the authors try to overcome this issue and study the estimators for $\beta-$mixing coefficients. Furthermore it has been investigated by Agarwahl and Duchi (2012) that in the setting when the underlying process is ergodic and either $\beta-$ or $\varphi-$mixing, performance of the online decision rules (used in the batch setting) is close to their regret. More precisely, if the loss-function is convex and the decision rule (after predicting the sequence $w_1, \ldots, w_n$ during time-steps $1 \leq t \leq n$) outputs the batch average of the predictors up to time n, it has small generalization error on future sample which is generated from the process with the same dependency condition.

In this work, in the framework of learning from dependent data observations we study the influence of (weak-)dependent data sample on the rates for non-parametric statistical learning schemes with values in reproducing kernel Hilbert spaces. In particular, we are interested in the high-probability control of the excess risk of the rule $f_{\mathcal{D}_n}$ which is returned by a general regularization procedure (1.7) based on the sample $\mathcal{D}_n$ of size n generated by some $\tau-$mixing process (see Chapter 2 for the definition of $\tau-$mixing). To be able to obtain these bounds we develop a general probabilistic toolbox of concentration inequalities of Bernstein-type for norms of partial sums of Banach-valued random variables. Under certain assumptions on the smoothness of the underlying regression function (posed in terms of the source

condition on the co-variance operator) and on the decay rate of the eigenvalues of the integral operator of the underlying RKHS (see section 4 in Chapter 2 or section 3 in Chapter 4 for more examples), we obtain that, in the case of geometrical decay of mixing coefficients (which is the case for example in the finite AR processes or in aperiodic recurrent Markov chains with finite state-space), up to a logarithmic factor the rates are the same as in the case with i.i.d. data observations.

1.3 Online (sequential) learning.

The second large part of this thesis is devoted to the problem of online (sequential) learning. We will mainly consider the so-called stochastic bandit problem (or sequential learning problem with partial feedback information) and the problem of adversarial online regression (sequential prediction with arbitrary data sequences). We give the mathematical preliminaries about the online learning problem. Let $\mathcal{X}$ be some convex topological set equipped with a Borel $\sigma-$algebra $\mathcal{B}(\mathcal{X})$. We refer to $\mathcal{X}$ as to the *instance* set, and let Θ be some topological space equipped with some $\sigma-$field $\mathcal{B}(\Theta)$ of the open sets. We refer to Θ as the parameter space or the *hypothesis space*. Consider also some abstract set of positive real measurable maps $\mathcal{L} = \{\ell : \Theta \mapsto \mathbb{R}_+\}$ which we refer to as to the set of losses. We present the general framework of online learning problems as a consecutive game between the *learner* and the *adversary* (or, in the framework with stochastic outputs, *environment*). The game unfolds as follows. At every round $t \geq 1$ a learner (optionally) observes some context $x_t \in \mathcal{X}$, and based on it and available feedback up to time t, chooses an element $\theta_t \in \Theta$. Next, the adversary chooses a loss-function $\ell_t \in \mathcal{L}$, the loss $\ell_t(\theta_t)$ is suffered and the game is repeated in the next round. The goal of the learner is to find a sequence of predictions $(\theta_t)_{t\geq 1}$ for which the cumulative loss $\sum_{t=1}^{T} \ell_t(\theta_t)$ will be small. To put mathematical details in this framework, we need to make precise what information is available to each player at every round of the game. We define the set which we call history available at time t to the learner as

$$\mathcal{H}_t = \{(x_s, h(\theta_s, \ell_s))_{s\leq t-1}\},$$

where $h : \Theta \times \mathcal{L} \mapsto \mathcal{G}$ is some measurable function, which we refer to as the feedback function and $(\mathcal{G}, \mathcal{B}(\mathcal{G}))$ is some measurable space. For a reason, which will be highlighted in Chapter 3 we also define the (extended) set in which the context information x_t at round t is available to the learner:

$$\mathcal{H}_t^o = \mathcal{H}_t \cup \{x_t\}$$

We say that the learning algorithm $\mathcal{A}$ which picks the sequence of decision rules $(\theta_t)_{t\geq 1}$ is admissible if for every t, the map $\mathcal{A}_t : (\mathcal{X} \times \Theta \times \mathcal{L})^{t-1} \mapsto \Theta$ is $(\sigma(\mathcal{H}_t), \mathcal{B}(\Theta))-$measurable. The goal of the last statement is to formalize the intuition that the learning algorithm cannot "see in the future". Furthermore, we assume that all contexts and losses $(x_t, \ell_t)_{t\geq 1}$ are fixed before the beginning of the game and do not depend on the choices of the algorithm $\mathcal{A}_t$ of the learner. Notice that this includes the stochastic framework as one can consider the sequence $\{x_s, \ell_s\}_{s\geq 1}$ as a trajectory of a random process with values in $\left(\mathcal{X}^{\mathbb{N}}, \mathcal{L}^{\mathbb{N}}\right)$.

Definition 1.3.1. We refer to the online learning problem as the problem with *full information* if h is the identity map from $\Theta \times \mathcal{L} \mapsto \Theta \times \mathcal{L}$, i.e. $h(\theta_s, \ell_s) = (\theta_s, \ell_s)$. We refer to the problem as a problem with bandit feedback if $h : \Theta \times \mathcal{L} \mapsto \mathbb{R}_+$ and $h(\theta_s, \ell_s) = \ell_s(\theta_s)$.

We describe separately the settings of online adversarial regression and the multi-armed stochastic bandit under the joint framework of learning from observed context x_t and feedback $\mathcal{H}_t$.

1.3.1 Online learning with full information. Adversarial online regression

In the framework of adversarial online learning, we consider the setting of sequential learning from arbitrary data sequences with full information and fixed loss function. First we provide some examples of learning frameworks with full information.

Example 1.3.2. Online binary classification. Let $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{0, 1\}$ and loss-function $\ell_t(\theta_t) := \ell(\theta_t, x_t, y_t) := \mathbb{I}_{\mathrm{Sgn}(\theta_t^\top x_t) \neq y_t}$. At each time $t \geq 1$ a learner obtains an instance $x_t \in \mathcal{X}$, makes a prediction $\widehat{y}_t = \mathrm{Sgn}(\langle \theta_t, x_t \rangle_{\mathbb{R}^d})$, where $\theta_t \in \Theta$ and $\Theta = B_1(\mathbf{0}) := \{\mathbf{x} \in \mathbb{R}^d : \|x\|_2 \leq 1\}$. Given the learner's prediction $\widehat{y}_t$, the adversary plays the loss function $\ell_t(x_t, y_t, \theta_t) := \ell(y_t, \widehat{y}_t) : \mathcal{X} \times \mathcal{Y} \times \widehat{\mathcal{Y}} \mapsto \mathbb{R}_+$ such that $\ell_t(, y_t, \widehat{y}_t) = \mathbb{I}_{y_t \neq \widehat{y}_t}$ (i.e. the learner suffers loss 1 when the prediction $\widehat{y}_t$ is wrong). A typical example is an online binary classification problem where the task is to correctly classify instances of 2 classes.

Example 1.3.3. Learning with expert advice. Let $\mathcal{X}$ be a subset of $\mathbb{R}^d$ and Θ an $n-$dimensional simplex, i.e. $\Theta = \{p \in [0, 1]^n : \|p\|_1 = 1\}$. Consider the linear loss function, i.e. $\ell_t(\theta_t) = \ell(x_t, \theta_t) = \langle x_t, \theta_t \rangle$. Then the regret is given as $R_T(\Theta) := \sum_{t=1}^T \langle p_t, x_t \rangle - \inf_{p \in \Theta} \sum_{t=1}^T \langle p, x_t \rangle$. This setting has a particular name, i.e. "learning with expert advice" which is natural as the prediction rule is essentially a linear combination of the "votes" of experts (which are represented as corners in the simplex).

An important algorithm which is used in the framework of learning with expert advice (but also can be extended to the general setting and learning with arbitrary loss functions) is the so-called exponentially weighted forecaster (see, e.g. in Littlestone and Warmuth (1994) for the introduction in the context of learning and see Cesa-Bianchi (1999a) for a broad survey).

Example 1.3.4. Online least-squares regression Let $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} := [-M, M]$ and $\theta \subset \mathcal{Y}^{\mathcal{X}}$, i.e. it is some subset of maps from $\mathcal{X}$ to $\mathcal{Y}$. Consider a least-squares loss-function $\ell_t : \mathcal{X} \times \mathcal{Y} \times \widehat{\mathcal{Y}} \mapsto \mathbb{R}$ $\ell(y_t, \widehat{y}_t) = (\widehat{y}_t - y_t)^2$, where the learner prediction is $\widehat{y}_t \in \mathcal{Y} \subset \mathbb{R}$, $\widehat{y}_t = \theta_t(x_t)$ and the environment outputs $y_t \in \mathcal{Y}$.

In this thesis (in Chapter 3) we will consider the problem of online sequential regression over arbitrary fixed data-sequences. First we specify the framework of online regression below. At the beginning of a round $t \geq 1$, learner observes a context $x_t \in \mathcal{X} \subset \mathbb{R}^d$ and makes a prediction $\widehat{y}_t := \theta_t(x_t)$ of the unknown label $y_t \in \mathcal{Y} \subset \mathbb{R}$, where element $\theta_t \in \Theta = \{\theta : \mathcal{X} \mapsto \mathbb{R}\}$ is the value of the algorithm $\mathcal{A}_t : \left(\mathcal{X} \times \mathcal{Y} \times \widehat{\mathcal{Y}}\right)^{t-1} \times \mathcal{X} \mapsto \Theta$ which is assumed to be $\sigma(\mathcal{H}_t)-$ measurable. Afterwards, the true label is revealed and loss $(y_t - \widehat{y}_t)^2$ is suffered. In this setting, at round $t \geq 1$ the learner has access to the history $\mathcal{H}_t = \sigma\{(x_s, \theta_s, y_s)_{s \leq t-1}\}$ The performance measure of the prediction rule is the loss against any fixed rule $\theta \in \Theta = \{\theta : \mathcal{X} \mapsto \mathbb{R}\}$ over all fixed sequences of pairs $\{x_t, y_t\}_{t=1}^T$, $T \in \mathbb{N}$. This leads to the notion of regret. We define the regret of admissible algorithm $\mathcal{A}$ which returns a sequence of predictions $(\widehat{y}_t)_{t=1}^T$ over data-sequence $\mathcal{D}_T = \{x_t, y_t\}_{t=1}^T$ as

$$R_T(\mathcal{A}, \mathcal{D}_T, \Theta) = \sup_{\theta \in \Theta} \left\{ \sum_{t=1}^T (y_t - \widehat{y}_t)^2 - \sum_{t=1}^T (y_t - \theta(x_t))^2 \right\}.$$

Because we focus on the complexity of the parameter space Θ, we use shorthand notation $R_T(\Theta)$ keeping dependence on $\mathcal{D}_T$ and $\mathcal{A}$ implicit. To characterize the notion of hardness of the problem, we introduce the concept of minimax regret over arbitrary fixed data-sequence $\mathcal{D}_T$, namely we introduce

$$\tilde{R}_T(\Theta) = \inf_{(\mathcal{A}_t)_{t \geq 1}} \sup_{\mathcal{D}_T = (x_t, y_t)_{t=1}^T} R_T(\Theta). \tag{1.14}$$

In the last quantity the infimum is taken over all admissible learning strategies $(\mathcal{A}_t)_{t \geq 1}$ with respect to the history $\mathcal{H}_t$, and the supremum over all fixed data-sequences $\mathcal{D}_T$ of size T with $x_t \in \mathcal{X}$, $y_t \in \mathcal{Y}$. We

say that algorithm $\mathcal{A}$ learns the best rule from the class Θ over any sequences of losses from class $\mathcal{L}$ if $\lim_{T\to\infty} \frac{R_T(\Theta,\mathcal{A},\mathcal{L})}{T} = 0$, and that the problem is online learnable over the loss-class $\mathcal{L}$ and parameter domain Θ. In Chapter 3 of this work, we aim to understand the behavior of a certain algorithm (a well-known nonparametric version of the Azoury-Warmuth-Vovk estimator) over some functional classes Θ of smooth functions (the so-called Sobolev classes) focusing on the compexity of the latter. Therefore, in the following we suppress dependence on $\mathcal{L}$ and $\mathcal{A}$ in the notation of the regret, keeping its dependence implicit and writing $R_T(\Theta)$.

Remark 1.3.5. First, notice that in this setting we do not pose any stochastic assumption on the data-sequence. Also notice that in the definition of the minimax regret, the supremum over all *fixed* data-sequences is used. Alternatively (since the learning algorithms are $\mathcal{H}_t-$measurable), we can write it by switching $\inf$ and $\sup$ and get (provided the output of the algorithm $\mathcal{A}$ on the step t is $\widehat{\theta}_t(x_t)$):

$$\tilde{R}_T(\Theta) = \inf_{\widehat{\theta}_1} \sup_{x_1,y_1} \ldots ... \inf_{\widehat{\theta}_T} \sup_{x_T,y_T} R_T(\Theta). \tag{1.15}$$

The origin of the online (sequential) prediction problem goes back to the work of Robbins (1952) where composed statistical problems have been considered. Firstly, sequential algorithms were considered in the works of Blackwell (1956), Hannan (1957). The very first studies of online models in the framework of learning from expert advice (a special case of learning with full information) are provided in the works by Littlestone and Warmuth (1994), Cesa-Bianchi et al. (1997) Vovk (1998). The game-theoretic setting from the perspective of online convex optimization was introduced by Zinkewich (2003), and since that developed in the thesesHazan (2006) and Shalev-Shwartz (2007). Further connections between online learning problems and, for example, finding equilibrium for economic systems are provided in Foster and Vohra (1997),Hart and Mas-Colell (2000). In the full-information framework, in the case when the decision set is finite (the so-called online learning with prediction of the expert advice), a survey on the forecaster and their regret analysis can be found in Cesa-Bianchi and Lugosi (2006).

The question of regret control in the problem of adversarial online regression is already interesting when Θ is a unit ℓ_2 norm bounded ball in the euclidean space $\mathbb{R}^d$. Here, the sequence of covariates $(x_t)_{t\geq 1}$ is the sequence of vectors in $\mathbb{R}^d$, and the decision rule is a linear model $w_t \in \Theta = \{w : \|w\|_2 \leq 1\}$, i.e. the prediction is $\widehat{y}_t = \langle w_t, x_t\rangle$. The (online) gradient descent algorithm which at time-step t computes $\theta_{t+1} = \theta_t - \eta(y_t - \langle\theta_t, x_t\rangle)x_t$ can be seen in the online framework as an application of a standard gradient descent iterative scheme to the square loss (see Kivinen and Warmuth (1997)). Furthermore, a standard and general approach in online learning is to apply a variation of the gradient iteration procedure, the so-called exponentiated gradient descent algorithm (see Kivinen and Warmuth (1997), Cesa-Bianchi (1999a)). It updates the coordinate i in the output vector by exponentially weighting the coordinates in which gradient's coordinates are far away from zero. Applying this general scheme to the squared-loss, the following update rule follows:

$$\theta_{t+1,i} = \frac{\theta_{t,i}\exp(-2\eta(y_t - \langle\theta_t, x_t\rangle)x_{t,i})}{\sum_{j=1}^d \theta_{t,j}\exp(-2\eta(y_t - \langle\theta_t, x_t\rangle)x_{t,j})}.$$

It can be shown that for bounded covariates x_t and $y_t \in [-B, B]$, the regret upper bound of the so-called "exponentially weighted average forecaster" (EWA) is of order $\left(4dB^2\ln(T)\right)$, which is optimal (with respect to the number of rounds T) on the classes of arbitrary balls in $\mathbb{R}^d$ and over the problem instances with bounded labels. An interesting observation is that this iterative update rule is equivalent (through performing online mirror descent update see Cesa-Bianchi (1999a)) to the following regularized optimization problem

$$\theta_t = \operatorname*{Arg\,Min}_{\theta\in\mathbb{R}^d} \left\|Y_{t-1} - X_{t-1}\theta^\top\right\|_2^2 + \lambda\|\theta\|_2^2,$$

where X_t is a $t-1 \times d$ matrix and $Y_{t-1} = (y_1, \ldots, y_{t-1})$ a vector of observed labels.

Furthermore, an improvement in terms of the multiplicative constant is achieved by the nonlinear ridge forecaster (see Azoury and Warmuth (2001), Vovk (2001) and also Gaillard et al. (2019)). Namely, it is shown that the so-called *Azoury-Warmuth-Vovk* algorithm achieves regret rates with improved leading constant of order $2B^2 d \ln(T)$ over the sets of bounded balls in $\mathbb{R}^d$ and furthermore this bound is optimal (see analysis for matching lower bound in Vovk (2001)) on the class of $d-$dimensional decision rules. According to this algorithmic scheme, the decision rule θ_t is the minimizer of the following optimization criterion:

$$\theta_t^{AWV} = \underset{\theta \in \mathbb{R}^d}{\text{Arg Min}} \left\| Y_{t-1} - X_{t-1}\theta^\top \right\|_2^2 + \lambda \|\theta\|_2^2 + \theta^2(x_t)$$

It can be shown that this algorithm is a variation of the so-called Aggregating Algorithm (see Vovk (1998)), which is used in the more general setting with convex losses. In the latter setting it enjoys an optimal regret of order $\sqrt{TK}$ (here K is the number of experts). Furthermore, in the setting of online linear regression the competing set Θ can be chosen to be the ball in ℓ_1 (Gerchinovitz and Yu (2013)), or the parameter space of the problem can change during the game (Hazan and Kale (2006)). The work Gerchinovitz (2013) deals with the problem of online linear regression with parameter set $\Theta = \{\theta : \theta = \sum_{j=1}^m \alpha_j \varphi_j\}$ for some given dictionary of vectors $\{\varphi_j\}_{j=1}^m$ in $\mathbb{R}^d$ and under some additional sparsity constraints (in the form of ℓ_1 or ℓ_0 norm) the minimizer over the set Θ; Langford et al. (2009) considers the ℓ_1- norm penalized regret in the linear case.

The setting becomes much broader when the underlying set Θ is some subset of some functional class of measurable functions, $\Theta \subset \widehat{\mathcal{Y}}^{\mathcal{X}}$. One standard idea in this setting is to discretize the space by means of some $\varepsilon-$net in L_∞ norm, obtaining in such a way a finite $\varepsilon-$ covering $\Theta_\varepsilon = \{f_1, \ldots, f_{K_\varepsilon}\}$ and using an exponentially weighted average forecaster on the finite set Θ_ε. In this case the learner is paying an approximation error ε over duration of the game T. This idea is introduced in Vovk (2006a), where it is applied to the broad classes of continuous functions over the balls of Besov and Sobolev spaces. It requires estimates of the metric entropy of the underlying balls. This idea can be further extended (see Gaillard and Gerchinovitz (2015)) to the case when instead of a fixed $\varepsilon-$net, one considers the successive refined approximations $\Theta_1, \Theta_2, \ldots, \Theta_{K_\varepsilon}$ of the underlying parameter space Θ and use an exponentially weighted average forecaster on the sets of successive differences between elements from consecutive sets Θ_i and Θ_{i+1}. Using the EWA on the series of consecutive approximations one obtains a "Chaining EWA" algorithm (see Gaillard and Gerchinovitz (2015)). It is proven to be optimal on the classes of bounded Hölder balls of smoothness level $\beta \in \mathbb{R}_+$.

Another way to characterize the complexity of online non-parametric regression problem is to characterize functional spaces by means of the convexity modulus of the unite sphere of the underlying Banach space (Clarkson (1936)). In this case, Vovk (2007) provides regret upper bounds in terms of the decay rate of modulus of convexity. When the competing space is the ball in the reproducing kernel Hilbert space, a strategy which is based on the approach of defensive forecasting (Vovk (2006b)) ensures regret of order $\mathcal{O}\left(\sqrt{T}\right)$.

Ideally, in the setting of online non-parametric learning, one would like to come up with an algorithm which has a small computational cost while being optimal in terms of regret rates (see Rakhlin and Sridharan (2014),Vovk (2001)).

In Chapter 3, we investigate the online variant of the non-parametric version of the Azoury-Vovk-Warmuth forecaster (also called Kernel averaging aggregation Regression or KAAR, see Gammerman et al. (2004)) with values in the RKHS whose kernel k is a Sobolev kernel (see Schaback (2007) for more information on Sobolev kernels). Namely, for some $s > \frac{d}{2}$ we consider the Sobolev RKHS $\mathcal{H}_k :=$

$W_2^s(\mathcal{X})$ and define the prediction rule as:

$$\widehat{f}_{t,\tau} = \underset{f\in\mathcal{H}_k}{\text{Arg Min}} \left(\|Y_{t-1} - S_{t-1}f\|_2^2 + \tau\|f\|_{\mathcal{H}_k}^2 + f^2(x_t) \right),$$

where $\tau > 0$ is some parameter and $S_t : \mathcal{H}_k \mapsto \mathbb{R}^t$ is the evaluation operator of function $f \in \mathcal{H}_k$ as before. One can easily find that $\widehat{f}_{t,\tau} = (S_t^* S_t + \tau\mathbb{I})^{-1} S_{t-1}^* Y_{t-1}$, where S_t^* is the correspondent adjoint to the operator S_t, as above we recall that it has the form $S_t^* Y_t = \sum_{i=1}^t y_s k_{x_s}$.

We show that the performance of the KAAR algorithm on the classes of Sobolev spaces $W_p^\beta(\mathcal{X})$ is essentially optimal when either $\beta > \frac{d}{2}$ and $p = \infty$ while having the least computational costs in comparison to the known algorithms which are based on the nested space discretization (see for example Gaillard and Gerchinovitz (2015)) or computationally not-tractable (see for example the seminal work of Rakhlin and Sridharan (2014)).

1.3.2 Stochastic bandits

Finally we consider the so-called (finite) multi-armed bandit (MAB) model. We first provide an informal description of the model as a game between the *environment* and the *learner* in the general setting as introduced above, and then give mathematical details. Namely, as described above, we consider the empty context set $\mathcal{X}$, the parameter set Θ to be the finite decision space, i.e. $\Theta = \{1, \dots, K\}$ and the feedback function is the evaluation of the loss ℓ_t over the action $\ell_t(a)$. The classical learner's goal is to minimize his loss, and the performance measure of the learner's strategy $(I_t)_{t\geq 1}$ is, as in the setting of online regression, a cumulative regret over $T \geq 1$ rounds with respect to any action $a \in \Theta$ which is fixed in hindsight:

$$R(a, I, T) = \sum_{t=1}^{T} \ell_t(I_t) - \ell_t(a). \tag{1.16}$$

Usually, in the setting of MAB, instead of losses one considers rewards and instead of a parameter set one refers to the action set. To simplify somewhat presentation the rewards are assumed to be bounded with values in $[0, 1]$. We consider also only bounded action sets and denote $\Theta = \mathcal{A} = \{1, \dots, K\}$ and denote the reward of action a in round t as the value X_t^a. Notice that for regret analysis in order to translate from the loss-setting to the reward setting one takes $X_t^a := 1 - \ell(a)$ and defines

$$R(a, I, T) = \sum_{t=1}^{T} (X_t^a - X_t^{I_t}). \tag{1.17}$$

The MAB problem of finding the best fixed action $a \in \mathcal{A}$ in hindsight demonstrates the so-called exploration-exploitation trade-off for the problem of regret minimization. Namely, in order to find the best action, an efficient algorithm has to explore between different actions thus increasing regret while not playing the optimal action. After having obtained some information about the environment, an efficient algorithm exploits it in order to minimize the regret.

The history of the bandit problem goes back to the work Thompson (1933), where William R. Thompson in the setting of performing medical trials was firstly to raise the question of sequential treatment allocation with adaptation "on the fly" to the drug, which at current moment appears to be the most effective. Statistical perspective of sequential experiment design was addressed in the work of Robbins (1952) where the action consisted in the choice of a prescribed treatment, and reward depends on the efficiency of its application for a patient. Since that work the field of sequential resources allocation has developed and substantially increased the number of domains where it can be of use. Namely, a variety of other different application problems can be modeled as a MAB problem. They include hyper-parameter

optimization, ad placement and applications in the computer games. We refer to the introductions in the surveys by Bubeck and Cesa-Bianchi (2012), Slivkins (2019) and recently Lattimore and Szepesvari (2020) for a broad overview of the possible spheres where the MAB can be applied.

Remark 1.3.6. We underline that the main challenge in the MAB problems is that the underlying rewards $(X_t^a)_{a\in\{K\}}$ are unknown and available to the learner only through the evaluation at single action $I_t \in \{K\}$. Furthermore, the notion of regret is defined for arbitrary sequences of (bounded) rewards without any stochastic assumptions. In this case one speaks of adversarial bandits (see Bubeck (2010), Cesa-Bianchi (1999a)). In this thesis, we however concentrate on the stochastic rewards scenario in which rewards $(X_t^a)_{t\in\mathbb{N},a\in\{K\}}$ are assumed to be generated from the real-valued stochastic processes over the space of all possible arm outcomes.

We introduce the setting of stochastic bandits more precisely. Let $\mathcal{A} = \{K\}$ be a finite action set; for some $a \in \{K\}$ consider an arbitrary probability space $(\Omega, \mathcal{F}, \mathbb{P}_a)$ and a stochastic process $(X_t^a)_{t\geq 1}$ with values in $\mathcal{X}^{\mathbb{N}} \subset [0,1]^{\mathbb{N}}$ defined over $(\Omega, \mathcal{F}, \mathbb{P}_a)$ is assumed to be stationary. Without loss of generality, we can always consider a canonical version of $(X_t^a)_{t\geq 1}$ over $(\mathcal{E}, \mathcal{B}(\mathcal{E}), \mu_a)$ where $\mathcal{E} = \mathcal{X}^{\mathbb{N}}$, $\mathcal{B}(\cdot)$-Borel $\sigma-$algebra and μ some measure over $\mathcal{E}$. We refer to a *bandit instance* as the joint distribution of the random process $(X_t^a)_{a\in\mathcal{A},t\in\mathbb{N}}$. More formally, for a set $\{K\}$ let a random process $B_t = \left(X_t^1, \dots, X_t^K\right)$ be defined over $(\Omega_B, \mathcal{A}, \mathbb{P}_{\mathcal{B}})$, where $\Omega_B = \left(\mathcal{X}^{\mathbb{N}}\right)^K$, $\mathcal{A} = \mathcal{B}\left(\mathcal{X}^{\mathbb{N}}\right)^{\otimes K}$ (i.e. the $\sigma-$algebra which is obtained by product of the cylinder sets) and $\mathbb{P}_{\mathcal{B}}$ is the probability distribution over $\mathcal{A}$ with $i-$th marginal $i \leq K$ being the distribution $\mathbb{P}_i$ of the process X_t^i over $\left(\mathcal{X}^{\mathbb{N}}, \mathcal{B}\left(\mathcal{X}^{\mathbb{N}}\right)\right)$. We call a Bandit instance independent if $\mathbb{P}^K$ is the product measure. Typical examples of bandit instances are independent Bernoulli bandits $\mathbb{P}_{Ber}^K = \nu_1 \otimes \nu_2 \otimes \dots \nu_K$, $\nu_i = B^{\otimes\mathbb{N}}(p_i)$, with $p_i \in [0,1]$; uniform bandits with $\mathbb{P}_{Uni}^K = U_1 \otimes \dots \otimes U_K$, $U_i = \mathcal{U}((a_i, b_i))^{\otimes\mathbb{N}}$, with $a_i, b_i \in [0,1]$ with $a_i \leq b_i$, subgaussian bandits $\mathbb{P}_{SG}^K = \mathbb{P}_1 \otimes \dots \otimes \mathbb{P}_K$ with $\mathbb{P}_i = \nu_i(\sigma_i)$ where $\nu_i(\sigma)$ is a subgaussian measure over $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with parameter $\sigma > 0$, i.e. such that $\int_{\mathbb{R}} e^{\lambda x}\nu_i(dx) \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right)$ for any $\lambda \in \mathbb{R}$. Notice that the case of subgaussian bandit instance in general violates the assumption that $(X_t^a)_{t\geq 1} \in [0,1]$, however the regret upper bounds (for the particular algorithms which are based on the confidence bounds analysis) are the same as in the bounded case, which is the subject to usage of subgaussian concentration inequality. For a bandit instance $\mathbb{P}_{\mathcal{B}}$, we will always assume (if the contrary is not stated) that all expectations and probabilities are taken under $\mathbb{P}_{\mathcal{B}}$ and suppress this in the notation. For $a \in \{K\}$ denote $\nu_a = \mathbb{E}[X_t^a]$, $a^* = \underset{a\in\mathcal{A}}{\text{Arg Max}}\, \mu_a$. We refer to the arm a^* as to the optimal arm, and for the arm $a \neq a^*$ we denote its suboptimality gap as the difference with the highest mean, namely we define $\Delta_a := \mu_* - \mu_a$. In the case of i.i.d. bandit instance, it becomes a natural measure of expected (single-round) regret suffered by the learner when pulling an arm $a \neq a^*$. Intuitively $(\Delta_a)_{a\in\{K\}}$ is a good characterization of the hardness of the bandit instance $\mathbb{P}_{\mathcal{D}_T}$, since if there are many $(\Delta_a)_{a\in\{K\}}$ with Δ_a being large, then the mean of μ_a is far from the mean of the optimal arm a_*. Therefore, it becomes easier (in the sense it requires less observations from each of the arms) to distinguish between them with given confidence level. Otherwise, if Δ_a is small then to distinguish between μ_a and μ_* one needs more samples, however in this case the difference between payoff of optimal arm and arm a is exactly Δ_a (thus being small). For the stochastic multi-armed bandit instance $\mathbb{P}_{\mathcal{B}}$, one typically differentiates between the notions of expected regret and pseudo regret. Namely, for any admissible strategy $(I_t)_{t\geq 1}$ its regret with respect to some action $a \in \{K\}$ over T rounds is defined as $R(a, I, T) = \sum_{t=1}^T \left(X_t^a - X_t^{I_t}\right)$. The expected regret under bandit instance $\mathbb{P}_{\mathcal{B}}$ is defined as the worst-case expected regret with respect to the best (fixed in hindsight) action under the bandit instance, i.e.

$$R_{\mathbb{P}_{\mathcal{B}}}(I, T) = \sup_{a\in\{K\}} \mathbb{E}_{\mathbb{P}_{\mathcal{B}}}[R(a, T)] = \max_{a\in\{K\}} \mathbb{E}_{\mathbb{P}_{\mathcal{B}}}[R(a, T)] = \mathbb{E}_{\mathbb{P}_{\mathcal{B}}}[R(\widetilde{a}, T)],$$

where $\widetilde{a} = \underset{a\in\{K\}}{\text{Arg Max}}\, \mathbb{E}_{\mathbb{P}_{\mathcal{B}}}[R(a, T)] = a^*$. In what follows we simplify the notation and implicitly assume the dependence of the regret on the learning strategy $(I_t)_{t\geq 1}$, thus focusing on the complexity of

expected regret in terms of bandit instance $\mathbb{P}_{\mathcal{B}}$ and number of rounds T. We define the pseudo-regret of a bandit instance $\mathbb{P}_{\mathcal{B}}$ as the following quantity:

$$\overline{R}_{\mathbb{P}_{\mathcal{B}}}(T) = \max_{a\in\{K\}} \mathbb{E}\big[\overline{R}(a,T)\big] := \max_{a\in\{K\}} \mathbb{E}\left[\sum_{t=1}^{T}(X_t^a - \mu_{I_t})\right],$$

which in case of stationary processes simplifies to

$$\overline{R}_{\mathbb{P}_{\mathcal{B}}}(T) = T\mu_* - \mathbb{E}_{\mathbb{P}_{\mathcal{B}}}\left[\sum_{t=1}^{T}\mu_{I_t}\right].$$

Notice that by Jensen's inequality notice it always holds that $\mathbb{E}\big[\max_{a\in\{K\}} R(a,I,T)\big] \geq R_{\mathbb{P}_{\mathcal{B}}}(I,T)$. Observe that when the bandit instance $\mathbb{P}_{\mathcal{B}} = \mathbb{P}_1 \otimes \ldots \otimes \mathbb{P}_K$ is a product measure, we have:

$$\begin{aligned} R_{\mathbb{P}_{\mathcal{B}}}(T) = T\mu_* - \mathbb{E}\left[\sum_{t=1}^{T} X_t^{I_t}\right] &= T\mu_* - \sum_{t=1}^{T}\sum_{k=1}^{K}\mathbb{E}\left[X_t^k \mathbb{I}_{I_t=k}\right] \\ &= \sum_{k=1}^{K}\mu_* \sum_{t=1}^{T}\mathbb{E}[\mathbb{I}_{I_t=k}] - \sum_{k=1}^{K}\sum_{t=1}^{T}\mathbb{E}\left[X_t^k\right]\mathbb{E}[\mathbb{I}_{I_t=k}] \\ &= \sum_{k=1}^{K}\Delta_k \mathbb{E}[N_k(T)], \end{aligned}$$

where $N_k(T) = \sum_{t=1}^{T}\mathbb{I}_{I_t=k}$, i.e. expected number of the times, the learner plays arm k and since $\sum_{t=1}^{T}\sum_{k=1}^{K}\mathbb{I}_{I_t=k} = T$. Furthermore, in this case (since random variable $X_t^{I_t}$ given the choice I_t is stochastically independent of I_t) we have $R_{\mathbb{P}_{\mathcal{B}}}(I,T) = \overline{R}_{\mathbb{P}_{\mathcal{B}}}(I,T)$. In such case the efficient algorithm (i.e. the one with low regret) will sample the sub-optimal arms (such that $\Delta_a > 0$) as small a number of times as possible. The stochastic bandit problem when $\mathbb{P}$ is a product measure was studied in various different settings. The celebrated UCB-algorithm (see Auer (2002)), which is based on the estimation of the upper confidence bound for the mean μ_a of arm a, was first mentioned in Lai (1987) . This procedure has the following regret upper bound (see Theorem 1 in Auer (2002)) for the stochastic payoff with values in $[0,1]$: $\overline{R}(T) \leq 8\sum_{i:\mu_a\leq\mu_*}\frac{\ln(n)}{\Delta_a} + \left(1+\frac{\pi^2}{3}\sum_{a\in\mathcal{A}}\Delta_a\right)$. We refer to the setting where the regret upper bound is determined in terms of gaps $(\Delta_a)_{a\in\mathcal{A}}$ as the problem-dependent bound. The term problem-dependent is justified as the inverse gaps $\frac{1}{\Delta_a}$ characterize the hardness of the problem. In this setting, it can be shown (see Lai and Robbins (1985) Theorem 1) that for arbitrary $\varepsilon > 0$ there exists no admissible algorithm such that for any bandit instance $\mathbb{P}_{\mathcal{B}}$

$$\overline{R}_{\mathbb{P}_{\mathcal{B}}}(T) \leq \sum_{a:\Delta_a>0}\frac{\log(n)}{(2+\varepsilon)\Delta_a},$$

uniformly over all distributions $\mathbb{P} = \mathbb{P}_1 \otimes \ldots \otimes \mathbb{P}_K$ with support in $[0,1]^K$. The latter bound means that UCB is the optimal policy in the case of problem-dependent regret bounds. In the case when the regret bounds are required to depend only on the number of arms K and time-horizon T we refer to them as to problem-independent bounds. In the setting of stochastic i.i.d. bandits it has been shown (see Auer (2002)) that UCB policy has regret upper bound of order $\sqrt{TK\log(T)}$, while the optimal lower bound is of order $\sqrt{TK}$ which is shown to be achieved by the so-called MOSS strategy Audibert and Bubeck (2009). A variation of the UCB algorithm (the so-called KL-UCB algorithm) is considered in Garivier and Cappé (2011). That algorithm enjoys uniformly better (in terms of multiplicative constant) regret upper bound than standard UCB while preserving it's asymptotic optimality on the class of bounded

independent bandit instances. Furthermore, a lot of different problems for a given bandit instance with a different performance measure (i.e. when instead of regret one has a different stochastic criterion based on the modeling assumption) have been considered in the literature. We mention only the problem of highest mean identification in a pure exploration framework (Even-Dar et al. (2002),Bubeck et al. (2009)), thresholding bandit problem (Locatelli et al. (2016)), and PAC-style problem (Evan-Dar et al. (2006)) where the goal is to find with probability at least $1-\delta$ any arm which is $\varepsilon-$close to the arm with the highest payoff.

In the setting with arm-dependent pay-offs, we consider the case when the distributions of the arms are bounded weakly-dependent $\phi_{\mathcal{C}}-$ mixing processes with a given dependence rate. Notice (see also discussion in Chapter 4) that in the case when the outputs of each arm are stochastically dependent over time, we have that $\mathbb{E}\left[\sum_{t=1}^{T} X_t^{I_t}\right] \neq \mathbb{E}\left[\sum_{t=1}^{T} \mu_t^{I_t}\right]$. In this situation an algorithm which minimizes the pseudo-regret (which means finding the arm with the highest stationary pay-off) can be different from an algorithm which minimizes the expected regret (which in this situation can be much smaller as the learner can exploit high dependencies between arm outcomes and sample from the arms when they output highly-correlated rewards). However, as, for example, shown in Grünewälder and Khaleghi (2017), fast decaying dependencies between past and the future (described by the so-called $\varphi-$ mixing coefficient therein) imply that the pseudo-regret can be a good proxy for the expected regret.

In Chapter 4 we analyse the pseudo-regret for a version of the so-called Improved-UCB Algorithm (see Ortner et al. (2014) and also Perchet and Rigollet (2013)) in the setting with weak-dependent bandit instances. We notice that in this scenario the usage of standard UCB-Algorithm (see Auer (2002)) is problematic, as the data samples collected in this scheme have strong couplings with respect to a first sample and the obtained process will not necessarily satisfy weak-dependency assumption (see discussion in Example 1 by Grünewälder and Khaleghi (2017)). The Improved-UCB does not have this disadvantage as the game timeline in this case is divided into epochs and in every epoch s the number of pulls and the pulls themselves are deterministic given the information to the beginning of the epoch s. We provide a broad analysis of both problem-dependent and problem-independent bounds for the pseudo-regret and describe the scenarios when (even in the case of slow decay correlations) the pseudo-regret bounds matches the scenario with i.i.d. arm's outcomes. In the case of strong correlations (the so-called slow-mixing scenario) we support the analysis with the (essentially) matching problem independent lower bound over the class of $\phi_{\mathcal{C}}-$mixing bandit instances.

1.4 General thesis overview

This thesis is organized as follows:

- Chapter 2 presents the results of the work "Concentration inequalities for weakly-dependent Banach-valued sums and applications to statistical learning methods". This is a joint work with Gilles Blanchard and is a published paper in Bernoulli 25 (4B) 3421 - 3458, November 2019. https://doi.org/10.3150/18-BEJ1095.

- Chapter 3 presents the results of the work "Online nonparametric regression with Sobolev kernels". This is a joint work with Pierre Gaillard, Sebastién Gerchinovitz and Alessandro Rudi and is avalaible as a preprint via https://arxiv.org/abs/2102.03594.

- Chapter 4 presents the results of the work "Restless dependent bandits with fading memory". This is a joint work with Gilles Blanchard and Alexandra Carpentier and is avalaible as a preprint via https://arxiv.org/abs/1906.10454.

- Chapter 5 presents the results of the work (under preliminary title) "Inequalities for dependent random fields". This is a joint work with Gilles Blanchard and Alexandra Carpentier which is an ongoing work in its finishing stage.

Each chapter is concluded by highlighting several potentially interesting scientific directions or problems which can be subject for the future work in the given field.

Chapter 2

Concentration of weakly-dependent random variables in Banach spaces

Contents

In this chapter we investigate the concentration of the sums of Banach-valued random variables which posses functional weak-dependency assumption of $\mathcal{C}-$mixing kind (see weak-dependency assumption (1.12)). We obtain a type of Bernstein inequality which is then used in the asymptotic framework to derive learning rates for the regularized learning methods based on a training sample from the $\tau-$mixing process. We discuss the sub-optimality of the obtained inequalities in certain cases and notice the interesting fact that, in the case of fast-decaying correlations (exponential weak-mixing), learning rates match (up to additional log-factor in the number of observations) the i.i.d. learning scenario.

This chapter of the thesis is based on the joint work with Gilles Blanchard, which can be found in Blanchard and Zadorozhnyi (2019).

2.1 Introduction

Let $(X_k)_{k\in\mathbb{N}_+}$ be an integrable and centered stochastic process taking values in a separable Banach space $(\mathcal{B}, \|\cdot\|)$. Define $S_n = X_1 + X_2 + \ldots + X_n$. We are interested in the non-asymptotic behaviour of the deviations of S_n from zero in $\mathcal{B}$; more precisely, we investigate exponential concentration inequalities for events of the type $\{\|S_n\| \geq t\}$, for $t > 0$. In the simplest situation where $(X_1, X_2, \ldots, X_n)$ are mutually independent and real-valued, the celebrated Hoeffding's Hoeffding (1963) and Bernstein's inequalities

Bernstein (1924) are available. Vector-valued analogues (in finite or infinite dimension) of those concentration inequalities for norms of sums of independent random variables were first established for the case of bounded independent random variables in Hilbert spaces by Yurinskyi Yurinskyi (1970).

However in an arbitrary Banach space the distribution of $\|S_n\|$ (in particular its expectation) heavily depends on the geometry of the underlying Banach norm. In this case moment (Bernstein-like) conditions for the individual variables X_i are not sufficient for the generic control of $\|S_n\|$ around zero (see Yurinskyi (1995), Example 3.0.1). Still, under assumptions on the "smoothness" of the underlying Banach norm (reflected by boundedness of its first two Gâteaux-derivatives), one can control the deviations of $\|S_n\|$ around zero. Corresponding concentration inequalities have been obtained in Pinelis and Sakhanenko (1986) and Pinelis (1992).

The case where random samples are generated from some stochastic process with possibly infinite memory are of interest for many applications. We review below some of the existing references on the topic of concentration of functions of random variables. The generalization of Hoeffding's inequality for real-valued martingales and martingale differences, together with its application to least squares estimators in linear and smooth autoregressive models are presented in van de Geer (2002). An extension of the Hoeffding-Azuma inequalities for the weighted sum of uniformly bounded martingale differences can be found in Rio (2013). Generalizations of the exponential inequalities for the case of real-valued supermartingales were obtained in Freedman (1975) and recently generalized in Fan et al. (2015), where the authors use change of probability measure techniques, and give applications for estimation in the general parametric (real-valued) autoregressive model. Extensions of Freedman (1975) for the case of supermartingales in Banach spaces were obtained in Pinelis (1994).

Beyond the (super)martingale setting, the need to handle more general processes which have some "asymptotic independence" assumptions led to the concept of mixing. Definitions of (strong) $\alpha-$, $\phi-$ and $\rho-$ mixing were introduced in Rosenblatt (1956), Ibragimov (1959) and Kolmogorov and Rozanov (1960), we refer also to Bradley (2005) for a broad survey about the properties and relations between strong mixing processes. However, there are examples of dynamical systems Dedecker et al. (2007) generated by uniformly expanding maps that are not $\alpha-$mixing (which is considered the weakest form of strong mixing assumptions, see Chapter 1 for comparison between the notions of mixing). Such types of processes include mixingales Andrews (1988); Mc Leish (1975), associated processes Esary et al. (1985); Fortuyn et al. (1971), and various more recent notions of weak dependence Bickel and Buehlmann (1999); Doukhan and Louhichi (1999); Rio (1996). We analyse the inherent dependency of the random sample by means of a functional weak-dependency assumption 1.12. In this setting, many techniques which were used in the independent data scenario were improved and combined with other methods to obtain concentration inequalities for the sums of *real-valued* random variables. Generalizations of Bernstein's inequality for ϕ-mixing random processes were obtained by making use of duality argument in the entropy method Samson (2000); using a blocking technique ensuring asymptotic independence, Bernstein-type inequalities for geometrically $\alpha-$mixing processes and moderate deviation principles were derived in Merlevede et al. (2009); deviation inequalities for real-valued sums of variables from general $\alpha-$mixing processes were obtained in Bosq (1993) through approximation by independent random sums and the blocking technique. Moreover, the blocking technique together with majorization of joint distributions by means of the marginals and a general Chernoff's bounding principle are used in Hang and Steinwart (2017) to obtain Bernstein-type inequalities for real-valued sums functions of $\mathcal{C}-$mixing processes (see Definition 1.2.3 in Chapter 1). In Kontorovich and Ramanan (2008), the martingale method is used to establish general McDiarmid-type concentration inequalities for real-valued Lipschitz functions of dependent random sequences on a countable state space. Using logarithmic Sobolev inequalities and the contractivity condition related to Dobrushin and Shlosman's strong mixing assumptions, general non-product measure concentration inequalities were obtained in Marton (2004).

Most of the above mentioned inequalities characterize the deviations of sums of real-valued random variables. Concerning Hilbert- or Banach-valued weakly dependent processes, a significant literature

exists on limit theorems of central limit or Berry-Esseen type, motivated in particular by functional time series Bosq (2000); Horváth and Kokoszka (2012). We limit ourselves to pointing out the recent reference Jirak (2018) and the substantial literature review there. This chapter is devoted to the concentration inequalities for the Banach norms of centered random sums with exponentially decaying deviation probability tails. A few results concern the concentration of real-valued functional of weakly dependent variables over general spaces, and can be applied to norms of sums of vector-valued variables. This is the case for the measure concentration inequalities due to Kontorovich and Ramanan (2008) for so-called $\eta-$mixing (which is implied by $\phi-$mixing) random variables, but a condition called Ψ-dominance Kontorovich (2006) must hold (it is satisfied if the underlying variable space is countable, or is a closed subspace of the real line). This result implies Hoeffding-Azuma type inequalities for norms of sums. Still, to the best of our knowledge, it is unknown how these mixing assumptions are connected to $\alpha-$, $\beta-$ or $\phi_{\mathcal{C}}-$mixing, or whether they can be applied to norms in arbitrary Banach spaces. The aforementioned measure concentration results of Marton (2004) for distributions of dependent real variables with continuous density imply concentration of the norm of their sum (which is a Lipschitz function in Euclidean distance) in an Euclidean space. However, the question becomes more challenging when one considers concentration of the norm of random variables in a separable, infinite-dimensional space. Finally, the recent work Dedecker and Merlevede (2015) establishes a Hoeffding-type bound under assumptions close to what we consider here; we underline that we are interested in sharper Bernstein-type rather than Hoeffding-type bounds (see also Section 2.3.3 for a more detailed discussion of the latter works).

This Chapter is constituted as follows: in Section 2.2, we recall the setting for stochastic processes with values in a Banach space. In Section 2.3, we pose the main assumptions about the structure of the underlying infinite dimensional Banach space and present, in a general form, the new Bernstein-type inequalities for $\mathcal{C}-$mixing processes. Furthermore, here we also provide specific corollaries for the cases of either exponentially (geometrically) or polynomially mixing decay rates. We compare our results to the former inequalities on the concentration of real-valued $\mathcal{C}$-mixing processes. In Section 2.4 we apply the obtained concentration inequalities in the statistical framework and analyze (in the asymptotic regime) the error bounds for reproducing kernel learning algorithms using a general form of spectral regularization when the sample is drawn from a process which satisfies the so-called τ-mixing assumption. All proofs can be found in the section 2.5.

2.2 Preliminaries and Notations

Let $(X_t)_{t\in\mathbb{N}}$ be a stochastic process defined over a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and valued in some bounded ball $\mathcal{X}$ of some separable Banach space $(\mathcal{B}, \|\cdot\|)$. Following Maume-Deschamps (2006), we give the definition of the weakly-dependent processes with respect to a class a of real-valued functions. Notice that in the form it is presented below, it is a particular case of dependence coefficient $a(\cdot,\cdot,\cdot)$, introduced in Equation (1.9) or (as an equivalent form) a particular case of projective-dependence coefficient $e_{\cdot,\cdot}(k)$ (see Chapter1). Let $\mathcal{M}_j = \sigma(X_i : 1 \leq i \leq j), j \in \mathbb{N}$.

Definition 2.2.1. For $k \in \mathbb{N}_{>0}$ we consider the $\mathcal{C}$-mixing coefficients as

$$\phi_{\mathcal{C}}(k) = \sup_{i\geq 1} \left\{ \mathrm{Cov}_{\mathbb{P}}[Y, \varphi(X_{i+k})], Y \in L_1(\Omega, \mathcal{F}_i, \mathbb{P}), \|Y\|_1 \leq 1, \varphi \in \mathcal{C}_1 \right\}.$$

We say that the process $(X_i)_{i\geq 1}$ is $\phi_{\mathcal{C}}-$mixing (or simply $\mathcal{C}$-mixing) if $\lim_{k\to\infty} \phi_{\mathcal{C}}(k) = 0$. If $\phi_{\mathcal{C}}(k) \leq c\exp(-bk^{\gamma})$ for some constants $b, \gamma > 0$, $c \geq 0$ and all $k \in \mathbb{N}$, then a stochastic process $(X_k)_{k\geq 1}$ is said to be *exponentially* (or *geometrically*) $\mathcal{C}$-mixing. If $\phi_{\mathcal{C}}(k) \leq ck^{-\gamma}$ for all $k \in \mathbb{N}$ and for some constants $c \geq 0, \gamma > 0$, then the stochastic process $(X_k)_{k\geq 1}$ is said to be *polynomially* $\mathcal{C}$-mixing.

As discussed in Maume-Deschamps (2006), $\mathcal{C}$-mixing describes many natural time-evolving systems

and finds its application for a variety of dynamical systems. Coefficients $\phi_{\mathcal{C}}$ are characterized by the control over correlations between the past and one moment in the future of the process over the class of bounded functions f such that $f \in \mathcal{C}_1$. An important result (Maume-Deschamps (2006), Lemma 1.1.2) claims that Definition 2.2.1 can be equivalently stated as follows:

Definition 2.2.2 (Equivalent to Definition 2.2.1)**.**

$$\phi_{\mathcal{C}}(k) = \sup_{i \geq 1} \left\{ \|\mathbb{E}[\varphi(X_{i+k})|\mathcal{M}_i] - \mathbb{E}[\varphi(X_{i+k})]\|_\infty \mid \varphi \in \mathcal{C}_1, \right\},$$

where $\|\cdot\|_\infty$ is the essential supremum norm as before. In our theoretical analysis we will use Definition 2.2.2 for processes which are assumed to be $\mathcal{C}-$mixing. We first give some examples of semi-norms C and thus provide particular types of dependency coefficients.

Example 2.2.3. Let $\mathcal{C}_{\text{Lip}}$ be the set of bounded Lipschitz functions over $\mathcal{X}$. Consider

$$C_{\text{Lip}}(f) := \|f\|_{\text{Lip}(\mathcal{X})} = \sup \left\{ \frac{|f(s) - f(t)|}{\|s - t\|} \;\middle|\; s, t \in \mathcal{X}, s \neq t \right\}.$$

It is easy to see that $C_{\text{Lip}}(f)$ is a semi-norm over set of bounded functions $\{f : \mathcal{X} \mapsto \mathbb{R}\}$. With this choice of class $\mathcal{C}$ and semi-norm $C(\cdot)$, we obtain the so-called $\tau-$mixing coefficients (see Dedecker et al. (2007) and Wintenberger (2010) for the real-valued case), which will be denoted $\tau(k) := \phi_{\mathcal{C}}(k), k \geq 1$.

Examples of $\tau-$mixing sequences. Consider a Banach-valued auto-regressive process of order 1

$$X_i = \rho(X_{i-1}) + \xi_i, \text{for } i \in \mathbb{Z},$$

where $(\xi_i)_{i\in\mathbb{Z}}$ is an i.i.d. sequence such that $\|\xi\| \leq 1$ almost surely, and $\rho : \mathcal{X} \mapsto \mathcal{X}$ is a linear operator with $\|\rho\|_* < 1$, where $\|\cdot\|_*$ is the operator norm. Due to the linearity of ρ, we can write $X_{t+s} = X_{t,s} + \rho^s(X_t)$, where $X_{t,s} = \sum_{l=0}^{s-1} \rho^l(\xi_{t+s-l})$. For the τ-mixing coefficients, by using this decomposition and the independence $X_{t,s}$ and X_t, we get:

$$\begin{aligned}
\tau(s) &= \sup_{f\in\mathcal{C}_1} \{\|\mathbb{E}[f(X_{t+s})|\mathcal{M}_t] - \mathbb{E}[f](X_{t+s})\|_\infty\} \\
&= \sup_{f\in\mathcal{C}_1} \{\|\mathbb{E}[f(X_{t,s} + \rho^s(X_t))|\mathcal{M}_t] - \mathbb{E}[f(X_{t,s} + \rho^s(X_t))]\|_\infty\} \\
&= \sup_{f\in\mathcal{C}_1} \{\|\mathbb{E}[f(X_{t,s} + \rho^s(X_t)) - f(X_{t,s})|\mathcal{M}_t] \\
&\qquad\qquad - \mathbb{E}[f(X_{t,s} + \rho^s(X_t)) - f(X_{t,s})]\|_\infty\} \\
&\leq 2\|\rho^s(X_t)\|_\infty \leq \|\rho\|_*^s \|X_t\|_\infty \to 0,
\end{aligned}$$

when $s \to \infty$, as X_t is almost surely bounded. From this we observe that $(X_t)_{t\geq 1}$ is exponentially $\tau-$mixing Banach-valued process. Repeating arguments from Andrews (1984) (in the real-valued case), one can show that this process is not always $\alpha-$mixing (in particular when ξ_i has a discrete distribution). Similarly as in the real case, it is easy to check that a Hilbert-valued version of the moving-average process of finite order $q < \infty$

$$W_i = \mu + \sum_{j=0}^{q} \theta_{i-j}\psi_{i-j}, \text{for } i \in \mathbb{Z},$$

where $(\psi_j)_{j\in\mathbb{Z}}$ is an independent and centered noise process and μ is some fixed element in a Hilbert space, is an exponentially $\tau-$mixing process. Furthermore, one can straightforwardly check that $(W_i)_{i\in\mathbb{Z}}$ is not a martingale in general.

Remark. We observe that the $\tau-$mixing property of the process $(X_t)_{t\geq 0}$ is preserved under a 1-Lipschitz map. More precisely, let $\phi : \mathcal{X} \mapsto \mathcal{H}$ be a $1-$Lipschitz mapping of the original process $(X_t)_{t\geq 0}$ to some Polish space $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$. Then, it is straightforward to check that the process $(\phi(X_t))_{t\geq 0}$ is again $\tau-$mixing. This conservation property is due to the definition of τ-mixing. The concentration inequality of Theorem 2.3.5 will allow us in Section 2.4 to obtain qualitative results about the statistical properties (error bounds) of the estimators of regression function in a reproducing kernel Hilbert space. The key idea here is that the estimators of the target function are based on a non-linear (but Lipschitz) mapping of the training data sequence into the Hilbert space where their they constitute a linear learning method.

Example 2.2.4. Assume $\mathcal{X} \subset \mathbb{R}$ to be an interval on the real line, let $\mathcal{C}_{\mathrm{BV}} := \mathrm{BV}(\mathcal{X})$ be the set of functions over $\mathcal{X}$ (bounded in total variation) and $C_{\mathrm{BV}}(\cdot)$ be the total variation seminorm:

$$C_{\mathrm{BV}}(f) := \|f\|_{\mathrm{TV}} = \sup_{(x_0,\ldots,x_n)\in\triangle} \sum_{i=1}^{n} |f(x_i) - f(x_{i-1})|,$$

where $\triangle = \{(x_0, x_1, \ldots, x_n) \in \mathcal{X}^n \mid x_0 < x_1 < \ldots < x_n\}$. It is known that $\mathrm{BV}(\mathcal{X})$ equipped with the norm $\|f\|_{\mathrm{BV}} = \|f\| + C_{\mathrm{BV}}(f)$ is a Banach space. With this choice of $(\mathcal{C}, C(\cdot))$ we obtain the so-called $\tilde{\phi}$-mixing processes, described in Rio (1996).

2.3 Main assumptions and results

2.3.1 Assumptions

Following Pinelis (1992), we introduce suitable hypotheses pertaining to the geometry of the underlying Banach space $(\mathcal{B}, \|\cdot\|)$, the distribution of the norm of coordinates $\|X_i\|$, and additional conditions on the considered $C(\cdot)$-semi-norm.

We recall briefly the concept of Gâteaux derivative: for a real-valued function $f : \mathcal{X} \to \mathbb{R}$ we say that f is *Gâteaux differentiable* at point $x \in int(\mathcal{X})$ in the direction $v \in \mathcal{B}$, if $t \mapsto f(x + tv)$ is differentiable in 0. We then denote

$$\delta_v f(x) = \left.\frac{d}{dt}\right|_{t=0} f(x + tv).$$

We say that the function f is *Gâteaux-differentiable* at point x if all the directional derivatives exist and form a bounded linear functional, i.e. an element $D_x f$ in the dual $\mathcal{B}^*$ such that $\forall v \in \mathcal{B}$:

$$\lim_{t\to 0} \frac{f(x + tv) - f(x)}{t} = \langle D_x f, v\rangle.$$

In this case $D_x f$ is called *Gâteaux derivative* of function f at point x.

Assumption A1. The norm $\|\cdot\|$ in the Banach space $\mathcal{B}$ is twice Gâteaux differentiable at every nonzero point in all directions and there exist constants $A_1 \geq 1, A_2 > 0$ such that the following conditions are fulfilled for all $x, v \in \mathcal{B}, x \neq 0$:

$$|\delta_v(\|x\|)| \leq A_1\|v\|, \text{ or equivalently } \left\|(D_x\|\cdot\|)\right\|_* \leq A_1;$$
$$|\delta_{v,v}(\|x\|)| \leq A_2\frac{\|v\|^2}{\|x\|},$$

where $\delta_{v,v}$ denotes the second Gâteaux differential in the direction v and $\|\cdot\|_*$ is the norm in the dual space $\mathcal{B}^*$. We give the following examples of Banach spaces that fulfill the **Assumption A1** (see Pinelis (1992) for the first two examples):

Example 2.3.1. Let $\mathcal{B} = \mathbb{H}$ be a separable infinite dimensional Hilbert space with scalar product $\langle \cdot, \cdot \rangle_{\mathbb{H}}$ and norm $\|\cdot\|_{\mathbb{H}}$. Then by triangle inequality, it holds:

$$\delta_g(\|f\|_{\mathbb{H}}) = \frac{d}{dt}\Big(\sqrt{\langle f+tg, f+tg\rangle}\Big)\Big|_{t=0} \leq \|g\|_{\mathbb{H}},$$

and also

$$\delta_{g,g}(\|f\|_{\mathbb{H}}) = \frac{d}{dt}\left(\frac{\langle f, g\rangle + t\|g\|^2_{\mathbb{H}}}{\|f+tg\|_{\mathbb{H}}}\right)\Big|_{t=0} \leq \frac{\|g\|^2_{\mathbb{H}}}{\|f\|_{\mathbb{H}}},$$

hence $\mathbb{H}$ satisfies Assumption **A1** with constants $A_1 = A_2 = 1$.

Example 2.3.2. Let $\mathcal{B} = L_p(\Omega, \mathcal{F}, \mathbb{P}), p \geq 2$. Then for any $f, g \in \mathcal{B}$ such that $f \neq 0$, it holds:

$$\begin{aligned}\delta_g(\|f\|_p) &= \frac{d}{dt}\left(\left(\int |f+tg|^p d\mathbb{P}\right)^{\frac{1}{p}}\right)\Big|_{t=0} = \|f\|_p^{1-p}\int |f|^{p-2} f g d\mathbb{P} \\ &\leq \|f\|_p^{1-p}\|f\|_p^{p-1}\|g\|_p = \|g\|_p,\end{aligned}$$

because of Hölder's inequality; similarly:

$$\begin{aligned}\delta_{g,g}(\|f\|_p) &= (p-1)\|f\|_p^{1-2p}\left(\|f\|_p\int |f|^{p-2}g^2 d\mathbb{P} - \left(\int |f|^{p-2} f g d\mathbb{P}\right)^2\right) \\ &\leq (p-1)\|f\|_p^{1-2p}\left(\|f\|_p\int |f|^{p-2}g^2 d\mathbb{P}\right) \leq (p-1)\|g\|_p^2\|f\|_p^{-1}.\end{aligned}$$

Thus for $p \geq 2$ an $L_p(\Omega, \mathcal{F}, \mathbb{P})$-space satisfies the conditions of Assumption **A1** with constants $A_1 = 1$, $A_2 = p-1$.

The next example characterizes the behaviour of the space of symmetric matrices equipped with a $p-$Shatten norm. It was not present in Pinelis (1992), is apparently new and we present it below for completeness.

Example 2.3.3. Let $p \in \mathbb{N}, p \geq 2$ be fixed and $\mathcal{B}_p$ be the space of real symmetric matrices of dimension d equipped with the Schatten $p-$norm $\|X\|_p = (\mathrm{Tr}(|X|^p))^{\frac{1}{p}} = \left(\sum_{i=1}^d |\lambda_i(X)|^p\right)^{\frac{1}{p}}$. Then it holds that for any elements $X, H \in \mathcal{B}_p$, $X \neq \mathbf{0}$:

$$\begin{aligned}\delta_H\big(\|X\|_p\big) &\leq \|H\|_p, \\ \delta_{H,H}\big(\|X\|_p\big) &\leq 3(p-1)\frac{\|H\|_p^2}{\|X\|_p},\end{aligned}$$

so the conditions of Assumption **A1** are satisfied with constants $A_1 = 1$ and $A_2 = 3(p-1)$ (for a detailed justification, see Appendix 2.3.1).

Proof of the Example 2.3.3 We make use of the following additional notation, which is standard in functional calculus over symmetric matrices. For a real diagonal matrix $W = \mathrm{diag}(w_1, \ldots, w_d)$ of dimension d, write $f(W) = \mathrm{diag}(f(w_1), \ldots, f(w_d))$, where $f : I \subset \mathbb{R} \to \mathbb{R}$ is a scalar function of class C^1 on I, and I is a finite union of open intervals of $\mathbb{R}$, containing the spectrum $\{w_1, \ldots, w_d\}$ of W.

For a symmetric matrix X with the spectral decomposition $X = U\Lambda U^\top = \sum_{i=1}^d \lambda_i e_i e_i^\top$ we consider the matrix-valued maps $f(X) = U f(\Lambda) U^\top$. Denote the real-valued function $g(X) = \mathrm{Tr}(f(X))$. Applying the chain rule and using Theorem V.3.3 from Bhatia (1997), we compute the Fréchet (and

hence Gâteaux) derivative of the function g at point X in the direction of an arbitrary matrix $H \in \mathcal{B}$. Namely, by linearity of the trace as a matrix operator, and from equations (V.9) and (V.12) from Bhatia (1997) (which are stated there in the case where I is an open interval, but the extension to a finite union of open intervals is immediate), we deduce:

$$\begin{aligned}\delta_H(g(X)) = \delta_H(\operatorname{Tr} f(X)) &= \frac{d}{dt}\Big|_{t=0} \operatorname{Tr} f(X+tH) = \operatorname{Tr} \frac{d}{dt}\Big|_{t=0} f(X+tH) \\ &= \operatorname{Tr} \delta_H f(X) = \operatorname{Tr}\left(f^{[1]}(\Lambda) \circ (U^\top H U)\right),\end{aligned}$$

where $\circ$ is used for the Hadamard (i.e. entry-wise) product of matrices; and $f^{[1]}(\Lambda)$ is a matrix whose (i,j) entry is defined as follows:

$$\left(f^{[1]}(\Lambda)\right)_{ij} = \begin{cases} \frac{f(\lambda_i)-f(\lambda_j)}{\lambda_i-\lambda_j}, & \text{if } \lambda_i \neq \lambda_j \\ f'(\lambda_i) & \text{otherwise.} \end{cases}$$

Thus, denoting $\widetilde{H} = U^\top H U$, we have:

$$\operatorname{Tr}\left(f^{[1]}(\Lambda) \circ \widetilde{H}\right) = \operatorname{Tr}\left(f'(\Lambda) \circ \widetilde{H}\right) = \operatorname{Tr}\left(f'(\Lambda)\widetilde{H}\right) = \operatorname{Tr}\left(f'(X)H\right) = \left\langle f'(X), H\right\rangle_F,$$

where the second to last equality follows from the definition of the matrix $f'(X)$ and $\langle \cdot, \cdot \rangle_F$ is the Frobenius product. This implies

$$\delta_H(\operatorname{Tr} f(X)) = \frac{d}{dt}\Big|_{t=0} \operatorname{Tr} f(X+tH) = \left\langle f'(X), H\right\rangle_F, \tag{2.1}$$

so that the Fréchet-derivative of $\operatorname{Tr}(f(X))$ is $f'(X)$. (This formula is certainly not a novelty and its justification included here for the sake of completeness.)

First consider the case where X has full rank, therefore has no zero eigenvalue, and apply Equation (2.1) to the function $f : t \mapsto |t|^p$ which is of class $\mathcal{C}^1$ on $I = \mathbb{R} \setminus \{0\}$, together with the chain rule to obtain that

$$\delta_H\left(\|X\|_p\right) = \frac{d}{dt}\Big|_{t=0} \|X+tH\|_p = \frac{d}{dt}\Big|_{t=0} \operatorname{Tr}(f(X+tH))^{\frac{1}{p}} = \left\langle \frac{w(X)}{\|X\|_p^{p-1}}, H \right\rangle_F, \tag{2.2}$$

where we introduced the notation $w(x) = \operatorname{sign}(x)|x|^{p-1}$ on I. From the definition of the Fréchet derivative, we have $D_X\left(\|\cdot\|_p\right) := \frac{w(X)}{\|X\|_p^{p-1}}$ is the corresponding Fréchet derivative at point X. Furthermore, for any $H \neq 0$ we have by the matrix variant of Hölder's inequality:

$$\frac{\delta_H\left(\|X\|_p\right)}{\|H\|_p} = \left\langle \frac{w(X)}{\|X\|_p^{p-1}}, \frac{H}{\|H\|_p} \right\rangle_F \leq 1,$$

thus, for any $H \in \mathcal{B}$ we have that $\left|\delta_H\left(\|X\|_p\right)\right| \leq A_1 \|H\|_p$ with constant $A_1 = 1$.

For the second Gâteaux differential, using linearity of the differential operator, we obtain:

$$\delta_{H,H}\left(\|\cdot\|_p\right) = \delta_H\left(\delta_H\left(\|\cdot\|_p\right)\right) = \delta_H\left(\left\langle D_X\left(\|\cdot\|_p\right), H\right\rangle\right) = \left\langle \delta_H\left(D_X\left(\|\cdot\|_p\right)\right), H\right\rangle.$$

Furthermore, for $\delta_H\left(D_X\left(\|\cdot\|_p\right)\right)$ we have by using the chain rule, (V.9) and (V.12) from Bhatia (1997)

again, differentiation rules for matrices and Equation (2.2):

$$\delta_H\big(D_X\big(\|\cdot\|_p\big)\big) = \frac{d}{dt}\Big|_{t=0} D_{X+tH}\big(\|\cdot\|_p\big) = \frac{d}{dt}\Big|_{t=0} \frac{w(X+tH)}{\|X+tH\|_p^{p-1}}$$
$$= \frac{U(w^{[1]}(\Lambda)\circ \widetilde{H})U^\top}{\|X\|_p^{p-1}} - (p-1)\frac{w(X)\langle w(X), H\rangle}{\|X\|_p^{2p-1}},$$

where the matrix $w^{[1]}(X)$ is defined analogously to $f^{[1]}(X)$ before. Therefore, for the second Gâteaux differential we obtain explicitly:

$$\delta_{H,H}\big(\|X\|_p\big) = \big\langle \delta_H\big(D_X\big(\|\cdot\|_p\big)\big), H\big\rangle = \frac{1}{\|X\|_p^{p-1}}\big\langle w^{[1]}(\Lambda)\circ \widetilde{H}, \widetilde{H}\big\rangle_F - (p-1)\frac{\langle w(X), H\rangle^2}{\|X\|_p^{2p-1}}. \quad (2.3)$$

For the second term, by the matricial Hölder's inequality we have:

$$(p-1)\frac{\langle w(X), H\rangle_F^2}{\|X\|_p^{2p-1}} \leq (p-1)\frac{\|X\|_p^{2p-2}\|H\|_p^2}{\|X\|_p^{2p-1}} = \frac{\|H\|_p^2}{\|X\|_p}.$$

For the first term, from the definition of the Hadamard product, we have

$$\Big\langle w^{[1]}(\Lambda)\circ \widetilde{H}, \widetilde{H}\Big\rangle_F = \sum_{i,j} w^{[1]}(\Lambda)_{ij}\widetilde{H}_{ij}^2.$$

Furthermore, taking into account that $w'(x) = (p-1)|x|^{p-2}$, by the mean value theorem on the closed interval $[\lambda_j, \lambda_i]$ (assuming $\lambda_i > \lambda_j$), since $p \geq 2$ the maximum of w' is attained at one of the endpoints of the interval, we have that

$$w^{[1]}(\Lambda)_{ij} \leq \frac{w(\lambda_i) - w(\lambda_j)}{\lambda_i - \lambda_j} \leq (p-1)\max\{|\lambda_i|^{p-2}, |\lambda_j|^{p-2}\}.$$

In the case where $\lambda_i = \lambda_j$, by definition $w^{[1]}(\Lambda)_{ij} = (p-1)|\lambda_i|^{p-2}$. Proceeding from this and using the symmetry of the matrix $\widetilde{H}$ after doing we get:

$$\sum_{i,j} w^{[1]}(\Lambda)_{ij}\widetilde{H}_{ij}^2 \leq \sum_{i,j}(p-1)\max\{|\lambda_i|^{p-2}, |\lambda_j|^{p-2}\}\widetilde{H}_{ij}^2$$
$$\leq (p-1)\sum_{i,j}\big(|\lambda_i|^{p-2} + |\lambda_j|^{p-2}\big)\widetilde{H}_{ij}^2$$
$$= 2(p-1)\sum_{i,j}|\lambda_i|^{p-2}\widetilde{H}_{ij}^2$$
$$= 2(p-1)\sum_i|\lambda_i|^{p-2}\sum_j \widetilde{H}_{ij}\widetilde{H}_{ji}$$
$$= 2(p-1)\sum_i|\lambda_i|^{p-2}\big(\widetilde{H}^2\big)_{ii}$$
$$= 2(p-1)\,\mathrm{Tr}\,\big(|\Lambda|^{p-2}\widetilde{H}^2\big).$$

Finally, applying the matricial Hölder's inequality once again for the last trace we get:

$$\mathrm{Tr}\Big(|\Lambda|^{p-2}\widetilde{H}^2\Big) = \Big\langle |X|^{p-2}, H^2\Big\rangle_F \leq \|X\|_p^{p-2}\|H\|_p^2.$$

Gathering the above estimates, for the first term of (2.3) we obtain the following bound:

$$\frac{1}{\|X\|_p^{p-1}}\Big\langle w^{[1]}(X)\circ\widetilde{H},\widetilde{H}\Big\rangle_F \leq 2\frac{(p-1)\|X\|_p^{p-2}\|H\|_p^2}{\|X\|_p^{p-1}} = 2(p-1)\frac{\|H\|_p^2}{\|X\|_p}.$$

The latter implies that $\big|\delta_{H,H}\big(\|X\|_p\big)\big| \leq A_2\frac{\|H\|_p^2}{\|X\|_p}$ with $A_2 = 3(p-1)$ for all $H \in \mathcal{B}$.

The inequalities required for Assumption **A1** are therefore established for all $X \in \mathcal{B}$ of full rank. To conclude the argument, it was established in Theorem 1 in Potapov and Sukochev (2014) that $X \mapsto \|X\|_p$ is of class $\mathcal{C}^{[p]}$ for all non-zero $X \in \mathcal{B}$. Since full rank matrices are dense in $\mathcal{B}$, $p \in \mathbb{N}$, by continuity, Assumption **A1** is satisfied for all non-zero $X \in \mathcal{B}$. □

The conditions in Assumption **A2** are common in the framework of Bernstein-type inequalities.

Assumption A2. There exist positive real constants c, σ^2 so that for all $i \in \mathbb{N}$:

$$\|X_i\| \leq c, \ \mathbb{P}\text{-almost surely};$$
$$\mathbb{E}\big[\|X_i\|^2\big] \leq \sigma^2.$$

Finally, being within the framework of the general weak-dependency assumption of Definition 2.2.2, we will consider functional classes $\mathcal{C}$ with a semi-norm $C(\cdot)$ which satisfies the following assumption.

Assumption A3. Let $C(f)$ be a semi-norm defined on a subspace $(\mathcal{C}, \|\cdot\|_{\mathcal{C}})$ of real bounded functions $\{f : \mathcal{X} \mapsto \mathbb{R}\}$. For every $s \in \mathcal{B}^*$ define $h_{1,s} : x \mapsto \langle s, x\rangle$ for each $s \in \mathcal{B}^*$ and $h_2 : x \mapsto \|x\|^2$, where $\mathcal{B}^*$ is the dual space of $\mathcal{B}$. Define $B(r)$, $B^*(r)$ to be the closed balls of radius r centered in zero in $\mathcal{B}$ and $\mathcal{B}^*$, respectively.

It is assumed that $h_{1,s} \in \mathcal{C}$ for all $s \in \mathcal{B}^*$; $h_2 \in \mathcal{C}$, and:

$$\sup_{s\in B^*(1)} C(h_{1,s}) \leq C_1,$$
$$C(h_2) \leq C_2,$$

for some fixed constants $C_1, C_2 \in \mathbb{R}_+$.

We write the constants C_1, C_2 from **Assumption A3** for the Examples 2.2.3, 2.2.4.

Example 2.2.3 (continued). For the Lipschitz class $\mathcal{C}_{\mathrm{Lip}}$ (see Example 2.2.3) we have:

$$\sup_{s\in B^*(1)} C_{\mathrm{Lip}}(h_{1,s}) = \sup_{s\in B^*(1)} \|h_{1,s}\|_{\mathrm{Lip}(B(c))} = \sup_{\substack{s\in B^*(1)\\ x_1,x_2\in B(c)}} \left\{\frac{\langle s, x_1 - x_2\rangle}{\|x_1 - x_2\|}\right\} \leq 1,$$

and

$$C_{\mathrm{Lip}}(h_2) = \|h_2\|_{\mathrm{Lip}(B(c))} = \sup_{x_1,x_2\in B(c)} \left\{\frac{\big|\|x_1\|^2 - \|x_2\|^2\big|}{\|x_1 - x_2\|}\right\} \leq 2c.$$

Example 2.2.4 (continued). For the BV functional class $\mathcal{C}_{\mathrm{BV}}$ considered in Example 2.2.4, and $\mathcal{X} = [-c, c] \subset \mathbb{R}$ we get (note that in this case $B^*(1) = [-1, 1]$ and the functional $h_{1,s}$ is just multiplication by s):

$$\sup_{s\in B^*(1)} C_{\mathrm{BV}}(h_{1,s}) = \sup_{|s|\leq 1} \|h_{1,s}\|_{\mathrm{BV}(B(c))} = \sup_{|s|\leq 1} \sup_{(x_0,\ldots,x_n)\in\triangle} \sum_{i=1}^{n} |s(x_i - x_{i-1})| = 2c.$$

$$C_{\mathrm{BV}}(h_2) = \|h_2\|_{\mathrm{BV}(B(c))} = \sup_{(x_0,\dots,x_n)\in\triangle} \sum_{i=1}^{n} |x_i^2 - x_{i-1}^2| \leq 2c^2.$$

2.3.2 Results

Our main result is a Bernstein-type inequality for norms of the sums of bounded Banach-valued centered $\phi_{\mathcal{C}}-$mixing random variables. We begin with a general bound on the deviations of the norm of $\sum_{i=1}^{n} X_i$.

Theorem 2.3.4. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be an arbitrary probability space, $(\mathcal{B}, \|\cdot\|)$ a Banach space such that Assumption* **A1** *holds and $\mathcal{X} = B(c)$. Let $(X_i)_{i\geq 1}^{n}$ be an $\mathcal{X}$-valued, centered, $\mathcal{C}$-mixing random process on $(\Omega, \mathcal{F}, \mathbb{P})$ such that Assumptions **A2, A3** are satisfied. Then for each pair of positive integers (ℓ, k), $\ell \geq 2$, such that $n = \ell k + r, r \in \{0, \cdots, k-1\}$, and any $\nu > 0$, it holds:*

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\right\| \geq 4A_1C_1\phi_{\mathcal{C}}(k) + 4\sqrt{\frac{B(\sigma^2 + C_2\phi_{\mathcal{C}}(k))\nu}{\ell}} + \frac{4c\nu}{3\ell}\right] \leq 2\exp(-\nu), \tag{2.4}$$

where $B = A_1^2 + A_2$ and the constants A_1, A_2, C_1, C_2 are given by the assumptions.

The choice of k and ℓ in the above result is related as $k = \lfloor \frac{n}{\ell} \rfloor$; thus one can optimize the obtained deviation bound over the choice of ℓ in order to reach the most favorable trade-off between the first term of order $\phi_{\mathcal{C}}(\lfloor \frac{n}{\ell} \rfloor)$ which is non-decreasing in ℓ, and the following "Bernstein-like" terms. This trade-off is a direct consequence of the blocking technique used in the proof of the above result. Namely the sample is divided into k blocks of size ℓ or $\ell + 1$, such that the distance between two neighbor points in the same block is exactly k. The Bernstein-like deviation terms are similar to the ones found in the i.i.d. case, with respect to sample size n replaced by the block size ℓ. The terms involving $\phi_{\mathcal{C}}$ reflect the lack of independence inside a block. This trade-off leads us to the notion of *effective sample size*. For a given n and constants c, σ^2 we define the positive integer number ℓ^*:

$$\ell^* := \max\left\{1 \leq \ell \leq n \text{ s.t. } C_1\phi_{\mathcal{C}}\left(\left\lfloor \frac{n}{\ell} \right\rfloor\right) \leq \frac{c}{\ell} \vee \frac{\sigma}{\sqrt{\ell}}\right\} \cup \{1\}. \tag{2.5}$$

Observe that ℓ^* is a function of n, but we omit this dependence to simplify notation. The following consequence of Theorem 2.3.4 is formulated in terms of the *effective sample size*:

Theorem 2.3.5. *Assume the conditions of Theorem 2.3.4 are satisfied, and the effective sample size ℓ^* is as given by (2.5). Then for any $\nu \geq 1$:*

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\right\| \geq \frac{\sigma(4A_1 + 6\sqrt{B}\sqrt{\nu})}{\sqrt{\ell^*}} + \frac{c(4A_1 + M_1\nu)}{\ell^*}\right] \leq 2\exp(-\nu), \tag{2.6}$$

where $M_1 := 2 + 2\sqrt{B}(1 + 2\frac{C_2}{C_1 c})$.

Remark. Lest the reader should wonder at the apparent lack of multiplicative scaling invariance in the last result due to the constant $C_2/(C_1 c)$ appearing in M_1, we stress that the $\mathcal{C}$-mixing assumption is not invariant with respect to the rescaling of the value space in general. However, in the particular cases of τ- and $\tilde{\phi}$-mixing (Examples 2.2.3, 2.2.4), the mixing assumption behaves gracefully with respect to scaling: in both cases it can be checked that the compound quantity $C_1\phi_{\mathcal{C}}(.)$ scales linearly with multiplicative rescaling of the space $\mathcal{X}$, such that the effective sample size ℓ^* given by (2.5) remains invariant, while $C_2/(C_1 c)$ remains constant, so that the deviation inequality (2.6) is unchanged by multiplicative rescaling, as one would expect.

Furthermore, we can give more explicit rates by lower bounding the effective sample size in the specific cases of exponentially or polynomially $\mathcal{C}-$mixing processes.

Proposition 2.3.6. *For an exponentially $\mathcal{C}-$mixing centered process on $(\Omega, \mathcal{F}, \mathbb{P})$ with rate $\phi_{\mathcal{C}}(k) := \chi \exp(-(\theta k)^{\gamma})$ ($\chi > 0, \theta > 0, \gamma > 0$), the effective sample size satisfies*

$$\ell^* \geq \left\lfloor \frac{n}{2} \theta \big(1 \vee \log\big(c^{-1} C_1 \chi \theta n\big)\big)^{-\frac{1}{\gamma}} \right\rfloor.$$

For a polynomially $\mathcal{C}-$mixing centered process with rate $\phi_{\mathcal{C}}(k) = \rho k^{-\gamma}$, the effective sample size satisfies

$$\ell^* \geq \max\left(\left\lfloor \left(\frac{\sigma}{C_1 \rho}\right)^{\frac{2}{2\gamma+1}} \left(\frac{n}{2}\right)^{\frac{2\gamma}{2\gamma+1}} \right\rfloor, \left\lfloor \left(\frac{c}{C_1 \rho}\right)^{\frac{1}{\gamma+1}} \left(\frac{n}{2}\right)^{\frac{\gamma}{\gamma+1}} \right\rfloor \right).$$

In the application section, we will use the obtained concentration framework for sums of Hilbert-space valued random variables. In this particular case, we have $A_1 = 1, A_2 = 1$ and correspondingly $B = 2$. Considering the case where the underlying data generating process is $\tau-$mixing (see Example 2.2.3) we get $C_1 = 1$ and $C_2 = 2c$. This gives us the following consequence for the concentration of the norm in the case of a process that satisfies the $\tau-$mixing conditions as in **Example 2.2.3.**

Corollary 2.3.7 (Concentration result for Hilbert-valued $\tau-$mixing processes)**.** *Under the assumptions of Theorem 2.3.5 with a Hilbert-valued $\tau-$mixing sample $\{X_i\}_{i=1}^n$, for any $0 \leq \eta \leq \frac{1}{2}$, with probability at least $1 - \eta$ it holds:*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} X_i \right\| \leq \log\left(\frac{2}{\eta}\right) \left(\frac{13\sigma}{\sqrt{\ell^*}} + \frac{21c}{\ell^*} \right), \tag{2.7}$$

where the choice of ℓ^* is given by (2.5).

2.3.3 Discussion

We highlight aspects in which our results differ from previous works. We first restrict our attention to the real-valued case ($\mathcal{B} = \mathbb{R}$). We consider the general type of $\phi_{\mathcal{C}}-$mixing processes as in Hang and Steinwart (2017), where the authors require the additional assumption on the semi-norm $C(\cdot)$ that the inequality $C(e^f) \leq \|f\|_{\infty} C(f)$ should hold for all $f \in \mathcal{C}$. Instead, we pose the assumption that the underlying class $\mathcal{C}$ contains linear forms and the function $x \mapsto \|x\|^2$, plus a.s. boundedness. The reason is that the proof of the main result essentially relies on the representation of the norm by means of its second order Taylor expansion. This allows us to recover results analogous to Hang and Steinwart (2017) (in the sense of the order of the effective sample size) for geometrically $\phi_{\mathcal{C}}-$mixing processes. As a broad overview and comparison to existing literature is given in Hang and Steinwart (2017); we omit reproducing this detailed discussion in the chapter and refer the reader to that work. As a further contribution with respect to Hang and Steinwart (2017), we derive new results for the exponential concentration of the sum for polynomially $\phi_{\mathcal{C}}-$mixing processes.

In the general Banach-valued case, the norm can be seen as a particular case of general functionals of the sample. As mentioned in the introduction to this chapter, while the literature on concentration of general functionals in the independent case is flourishing, it is rather scarce for the setting of weak dependence. In the work Kontorovich and Ramanan (2008), the authors obtain general Hoeffding-type concentration inequalities for functionals of the sample satisfying the bounded difference assumption (Azuma-McDiarmid type setting) under the so-called $\eta-$mixing assumption (which is related to, but weaker than, ϕ-mixing). The core proof technique in our results as well as in Kontorovich and Ramanan (2008) is the martingale difference approach.

Furthermore, in the work Dedecker and Merlevede (2015), the authors establish a MarcinkiewiczZygmund type inequality for dependent random Banach-valued sums under assumptions on the smoothness of the corresponding norm which are very close to ours. In particular, from Corollary 3.2 in Dedecker and Merlevede (2015), one can deduce that for a bounded $\tau-$mixing process $(X_i)_{i \geq 0}$ (see **Example 2.2.3.**)

with values in $\mathbb{L}^q(\Omega, \mathcal{A}, \mathbb{P})$ for $q \geq 2$, a Hoeffding-type exponential bound holds for sums with a deviation rate of order $2cb_n\sqrt{\log(e/\delta)/n}$, where c is as in Assumption **A2**, and $b_n^2 := 1 + \sum_{i=1}^n \tau(i)$.

Comparing to this last deviation bound, an advantage of our results is that they are of Bernstein- rather than Hoeffding-type, and valid under a weaker dependence assumption, which includes $\tau-$mixing as specific case. On the other hand, the deviation scaling in b_n above is better than ours (this is relevant for polynomial mixing conditions). This leaves open for future work the question of scaling in the framework of Bernstein-type inequalities under the assumptions we consider and their improvement in terms of dependency on the mixing rate.

The strongest assumption we make (besides those concerning the geometry of $\mathcal{B}$ and the class $\mathcal{C}$) is the a.s. boundedness of the random variable; this assumption was also considered in Pinelis and Sakhanenko (1986) and Yurinskyi (1970) (for Bernstein-type inequalities for a Banach-valued independent process, which is included in the present result) and in Hang and Steinwart (2017) (for Bernstein-type inequality with weakly dependent real variables). From a technical point of view, our current proof relies significantly on that assumption at several key places; removing this assumption to replace it with a weaker control of moments (as in the classical independent real-valued Bernstein inequality) is a stimulating question.

We now apply the concentration results to the particular case of random variables with values in a separable Hilbert space, and use them for the analysis of statistical properties of kernel-based algorithms in machine learning which are trained on a dependent sample. This analysis will be the cornerstone of the next section.

Notice however, that using the methods developed in Chapter 5 for the case of random fields the results can be extended to obtain sharper inequalities (in terms of deviation rates) in the case of Banach-valued random sums.

2.4 Application to statistical learning

Let $\mathcal{X}$ be a closed ball of a Polish space and $\mathcal{Y} = \mathbb{R}$. Consider a stationary stochastic process $(Z_i)_{i\geq 1}$ over some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $\mathcal{X} \times \mathcal{Y}$, and define ν as the common marginal distribution of the Z_is, and μ as its X-marginal. We denote $\nu(\cdot|\cdot)$ a regular conditional probability distribution of Y_i conditional to X_i. In the general framework of learning from examples, the goal is to find a prediction function $f : \mathcal{X} \mapsto \mathcal{Y}$ such that for a new pair $(X, Y) \sim \nu$, the value $f(X)$ is a good predictor for Y. Let $\mathcal{D}_n := \{x_i, y_i\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ be the observed training sample from the n first coordinates of the process $(Z_i)_{i\geq 1}$, and $f_{\mathcal{D}_n}$ be an estimated prediction function belonging to some model class $\mathcal{H}$. We will assume $(Z_i)_{i\geq 1}$ to be a $\tau-$mixing stationary process (as in Example 2.2.3) on $(\Omega, \mathcal{F}, \mathbb{P})$. We consider the least squares regression problem where the goal is to minimize the averaged squared loss $R_{LS,\mu}(f) := \mathbb{E}_\nu\big[(f(X) - Y)^2\big]$. Equivalently, we want to find $f_{\mathcal{D}_n}$ that approximates the regression function $f_\nu(x) = \mathbb{E}[Y|X = x]$ well in the sense of being close to optimal risk $\mathcal{E}(f)$ over the considered model class.

2.4.1 Learning by means of reproducing kernels

We investigate statistical learning methods based on reproducing kernel Hilbert space regularization. As a set of decision rules we consider a separable real reproducing kernel Hilbert space (RKHS) $\mathcal{H} = \mathcal{H}_k \subset L^2(X, \mu)$ which is induced by a measurable kernel k over $\mathcal{X}^2$. An in depth survey on the kernel methods can be found in Steinwart and Christmann (2008), Smola and Schölkopf (2002), see also introduction to the setting of learning with kernels in Chapter 1.

Reproducing kernel Hilbert spaces are of broad usage in the non-parametric learning in particular because of the linear structure of the solutions to many optimization problems. In the next pages we recall the setting and notation used in the framework of statistical learning; more details can be found in

Bauer et al. (2009); Blanchard and Mücke (2018) (for the inverse learning perspective); also in Caponetto and De Vito (2007); Rosasco et al. (2010) (for the statistical learning perspective).

In our analysis we assume the kernel to be bounded by a positive constant $\kappa = 1$, i.e. $\sup_{x\in\mathcal{X}} \sqrt{k(x,x)} \leq 1$. This implies that any $f \in \mathcal{H}_k$ is measurable and bounded in the supremum norm. As $\mathcal{H}_k$ is a subset of $L^2(\mathcal{X},\mu)$, let $S_k : \mathcal{H}_k \mapsto L^2(\mathcal{X},\mu)$ be the inclusion operator; and $S_k^* : L^2(\mathcal{X},\mu) \mapsto \mathcal{H}_k$ its adjoint. Analogously as in Chapter 1 we define operators T, L and the empirical counterparts $S_n, S_n, T_n, L_n := K_n$.

We now specify classes of distributions which correspond to a certain regularity of the learning problem in relation to the RKHS $\mathcal{H}_k$, and on which we establish the error bounds. We start with the following assumption on the underlying distribution ν and the corresponding regression function f_ν.

Assumption B1. There exist $0 < R \leq 1, \Sigma > 0$ such that the distribution ν belongs to the set $\mathcal{D}(R,\Sigma)$ of distributions satisfying:

i) $|Y| \leq R$, ν-almost surely.

ii) The regression function f_ν belongs to the RKHS $\mathcal{H}_k$, i.e. for μ-almost all $x \in \mathcal{X}$ it holds

$$\mathbb{E}\big[Y|X=x\big] := \int_{y\in\mathcal{Y}} y\nu(dy|x) = f_\nu(x),\ f_\nu \in \mathcal{H}_k.$$

iii) For $\mu-$almost all x:

$$\mathrm{Var}\big(Y|X=x\big) = \int_{y\in\mathcal{Y}} (y - f_\nu(x))^2 \nu(dy|x) \leq \Sigma^2.$$

Point (i) of the assumption ensures that we can assume $\mathcal{Y} = [-R, R]$ without loss of generality. The two next assumptions are: a decay rate condition for the discrete spectrum $(\zeta_i)_{i\geq 1}$ (ordered in decreasing order) of the covariance operator T, and the so-called Hölder source condition (see e.g. De Vito et al. (2006)) that describes the smoothness of the regression function f_ν. Denoting $\mathcal{P}$ to be the set of all probability distributions on $\mathcal{X}$; we will thus assume that the X-marginal distribution μ belongs to

$$\mathcal{P}^<(b,\beta) := \big\{\mu \in \mathcal{P} : \zeta_j \leq \beta j^{-b}, \forall j \geq 1\big\};$$

secondly, we assume that $f_\nu \in \Omega(r,D)$, where

$$\Omega(r,D) = \big\{f \in \mathcal{H}_k | f = T^r g, \|g\|_{\mathcal{H}_k} \leq D\big\}, \tag{2.8}$$

for some $r \geq 0$, which in the inverse problems literature is called the standard Hölder source condition for the linear embedding problem. Joining all assumptions, we consider the following class of marginal generating distributions:

$$\mathcal{M}(R,\Sigma,r,D,\beta,b) := \big\{\nu(dx,dy) = \nu(dy|x)\mu(dx) : \nu \in \mathcal{D}(R,\Sigma), \mu \in \mathcal{P}^<(b,\beta), f_\nu \in \Omega(r,D)\big\}. \tag{2.9}$$

For estimation of the target regression function f_ν, we consider the following class of kernel spectral regularization methods:

$$f_{\mathcal{D}_n}^\lambda = F_\lambda(T_n) S_n^* \mathbf{y}, \tag{2.10}$$

where $F_\lambda : [0,1] \mapsto \mathbb{R}$ is a family of functions. The expression $F_\lambda(T_n)$ is to be understood in the usual sense of (compact, self-adjoint) functional calculus on operators. The family $(F_\lambda)_{\lambda\in[0,1]}$ defines the regularization method (which we also call regularization function), depending on the parameter $\lambda \in (0,1]$, and for which the following conditions hold:

i) There exists a constant $B < \infty$ such that, for any $0 < \lambda \leq 1$:

$$\sup_{t\in(0,1]} \left|tF_\lambda(t)\right| \leq B.$$

ii) There exists a constant $E < \infty$ such that

$$\sup_{t\in(0,1]} \left|tF_\lambda(t)\right| \leq E/\lambda.$$

iii) There exists a constant γ_0 such that the *residual* $r_\lambda(t) := 1 - F_\lambda(t)t$ is uniformly bounded, i.e.

$$\sup_{t\in(0,1]} \left|r_\lambda(t)\right| \leq \gamma_0.$$

iv) For some positive constant γ_q there exists a maximal q, which is called *the qualification of the regularization* such that

$$\sup_{t\in(0,1]} \left|r_\lambda(t)t^q\right| \leq \gamma_q\lambda^q.$$

The above conditions are standard in the framework of inverse problems and in an asymptotic framework are sufficient (see Bauer et al. (2009)) in order to obtain consistent learning algorithms in the case of independent examples. Many known regularization procedures (including Tikhonov regularization, spectral cut-off, Landweber iteration) may be obtained as special cases via the appropriate choice of the regularization function F_λ and satisfy conditions $i)-iv)$ for appropriate parameters. In the work by Engl et al. (1996), Bauer et al. (2009) and Rosasco et al. (2010) a variety of different examples of regularization learning methods as well as the discussion in the context of learning from independent examples are provided.

2.4.2 Learning from a $\tau-$mixing sample

For the learning part we restrict to the case of $\tau-$mixing processes, see Example 2.2.3.

Obtaining probabilistic results for Hilbert-valued estimators (analogous in spirit to those in Blanchard and Mücke (2018)), we derive upper bounds on the estimation error of f_ν by regularized kernel learning estimators $f^\lambda_{\mathcal{D}_n}$; in the case of learning from τ-mixing samples, assuming a polynomial spectrum decay rate of the covariance operator T, and for a certain range of norms.

A key technical tool used in previous works for the analysis of the i.i.d. case (see Bauer et al. (2009) and Blanchard and Mücke (2018)) is a quantitative statement for the concentration of the centered (and possibly suitably rescaled) Hilbert-space valued variables $(S^*_{\mathbf{x}}\mathbf{y} - T_{\mathbf{x}}f_\nu)$ and $(T_n - T)$ around 0. Observe that these variables are empirical sums (of elements $k_{x_i}(y_i - f_\nu(x_i)) \in \mathcal{H}_k$ and $(k_{x_i} \otimes k^*_{x_i} - T) \in \mathrm{HS}(\mathcal{H}_k)$, respectively). Thus, a very natural way to proceed in the analysis is to use the concentration results established in Section 2.3 for Hilbert spaces as replacement for their i.i.d. analogues, and for other steps to follow the proof strategy of those earlier works.

Assuming the sample $\mathcal{D}_n = \{x_i, y_i\}_{i=1}^n$ is a realization from a $\tau-$mixing process $(Z_i)_{j\geq 1}$, in order to apply the concentration inequality from Section 2.3, we should ensure that the corresponding Hilbert-valued quantities are forming a τ-mixing sequence themselves. As pointed out earlier, the τ-mixing property is obviously preserved (up to constant) via a Lipschitz map. Lemma 2.6.1 establishes this Lipschitz property for the kernel maps under mild assumptions (uniformly bounded mixed second derivative of the kernel). Using the inequality from Corollary 2.3.7, in Lemma 2.4.1 we obtain high probability inequalities for deviations of the corresponding random elements. The proof of the lemma can be found in Appendix 2.6. To simplify the exposition, we specify the results for the cases of either exponentially or polynomially mixing process. Further extensions are possible using the same general proof scheme as

a blueprint, described in Appendix 2.6, together with the result of Theorem 2.3.5 on the effective sample size.

Lemma 2.4.1. *Let $\mathcal{X}$, $\mathcal{Y} = [-R, R]$ and $\mathcal{H}_k$ be as defined before. Assume that the kernel k satisfies $\sup_{x\in\mathcal{X}} \sqrt{k(x,x)} \leq 1$ and admits a mixed partial derivative $\partial_{1,2}k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ which is uniformly bounded by some positive constant K. Let $(Z_j = (X_j, Y_j))_{j\geq 1}$ be a $\tau-$mixing process with rate $\tau(k)$, satisfying Assumption* **B1** *and such that $\|f_\nu\| \leq D$.*

For any $\eta \in (0, 1/2]$ the probability of each one of the following events is at least $1 - \eta$:

$$
\begin{aligned}
\left\|T_n f_\nu - S_n^* y\right\|_{\mathcal{H}_k} &\leq 21 \log\left(2\eta^{-1}\right)\left(\frac{\Sigma}{\sqrt{\ell_1}} + \frac{2R}{\ell_1}\right);\\
\left\|(T+\lambda)^{-\frac{1}{2}}(T_n f_\nu - S_n^* y)\right\|_{\mathcal{H}_k} &\leq 21 \log\left(2\eta^{-1}\right)\left(\frac{\Sigma\sqrt{\mathcal{N}(\lambda)}}{\sqrt{\ell_2}} + \frac{2R}{\sqrt{\lambda}\ell_2}\right);\\
\left\|(T+\lambda)^{-1/2}(T - T_n)\right\|_{\mathrm{HS}(\mathcal{H}_k)} &\leq 21 \log\left(2\eta^{-1}\right)\left(\frac{\sqrt{\mathcal{N}(\lambda)}}{\sqrt{\ell_3}} + \frac{2}{\sqrt{\lambda}\ell_3}\right);\\
\left\|T - T_n\right\|_{\mathrm{HS}(\mathcal{H}_k)} &\leq 42\frac{\log\left(2\eta^{-1}\right)}{\sqrt{\ell_4}},
\end{aligned}
\tag{2.11}
$$

where the quantity $\mathcal{N}(\lambda) := Tr\left((T+\lambda)^{-1}T\right)$ is the so-called effective dimension*; $\ell_1, \ell_2, \ell_3, \ell_4$ are in each case suitable bounds on the effective sample size. For exponentially and polynomially $\tau-$mixing rates, corresponding bounds for effective sample sizes are given in Table 2.1.*

Table 2.1: Bounds on effective samples sizes for (2.11). Here we set $C := 3\max(1, KR, KD)$.

	$\tau(k) = \chi\exp(-(\theta k)^\gamma)$	$\tau(k) = \rho k^{-\gamma}$
ℓ_1	$\left\lfloor \frac{n\theta}{2\left(1\vee\log\left(n\frac{C\chi\theta}{2R}\right)\right)^{\frac{1}{\gamma}}} \right\rfloor$	$\left\lfloor \left(\frac{\Sigma}{C\rho}\right)^{\frac{2}{2\gamma+1}}\left(\frac{n}{2}\right)^{\frac{2\gamma}{2\gamma+1}} \right\rfloor$
ℓ_2		$\left\lfloor \left(\frac{\Sigma\sqrt{\lambda\mathcal{N}(\lambda)}}{C\rho}\right)^{\frac{2}{2\gamma+1}}\left(\frac{n}{2}\right)^{\frac{2\gamma}{2\gamma+1}} \right\rfloor$
ℓ_3	$\left\lfloor \frac{n\theta}{2(1\vee\log(nK\theta\chi))^{\frac{1}{\gamma}}} \right\rfloor$	$\left\lfloor \left(\frac{\sqrt{\lambda\mathcal{N}(\lambda)}}{2K\rho}\right)^{\frac{2}{2\gamma+1}}\left(\frac{n}{2}\right)^{\frac{2\gamma}{2\gamma+1}} \right\rfloor$
ℓ_4		$\left\lfloor \left(\frac{1}{K\rho}\right)^{\frac{2}{2\gamma+1}}\left(\frac{n}{2}\right)^{\frac{2\gamma}{2\gamma+1}} \right\rfloor$

Remark 2.4.2. The first inequality will not be used in the statistical analysis to follow and is presented here for completeness. We notice also that the effective dimension is the key quantity in the risk analysis of the Hilbert-valued regularization scheme. Since operators L and T have the same spectrum, one can write $\mathcal{N}(\lambda) = \sum_{j\geq 1}\frac{\lambda_j(T)}{\lambda_j(T)+\lambda} = \sum_{j\geq 1}\frac{\lambda_j(L)}{\lambda_j(L)+\lambda}$ and the analysis of $\mathcal{N}(\lambda)$ boils down to the analysis of the eigenvalues (in particular their decay rates) of a kernel integral operator.

Armed with the above probabilistic results, we derive upper bounds for the errors of estimation of f_ν by means of the general regularized kernel learning estimators (2.10). The main tool is the following lemma, giving a high probability inequality on the deviation of the estimation error. The gist of this result and of its proof is to follow the approach of Blanchard and Mücke (2018), wherein the sample size in the i.i.d. case is replaced by the effective sample size, the rest of the argument being essentially the same.

Lemma 2.4.3. *Consider the same assumptions as in Lemma 2.4.1. Assume that $f_\nu \in \Omega(r, D)$ (defined by (2.8)) for some positive numbers r, D. Also, let $f_{\mathcal{D}_n}^\lambda$ be the regularized estimator as in (2.10), with*

a regularization satisfying conditions (i)-(iv) with qualification $q \geq r+s$. Fix numbers $\eta \in (0,1]$ and $\lambda \in (0,1]$ and denote:

$$\overline{\gamma} := \max(\gamma_0, \gamma_q), \qquad \ell_0 := 2500\lambda^{-1}\max(\mathcal{N}(\lambda),1)\log^2\left(\frac{8}{\eta}\right),$$

where we recall that γ_0, γ_q are the constants from conditions iii)-iv).

Then with probability at least $1-\eta$, the inequality

$$\left\|T^s\left(f_{\mathcal{H}_k} - f^{\lambda}_{\mathcal{D}_n}\right)\right\|_{\mathcal{H}_k} \leq C_{r,s,B,E,\overline{\gamma}}\log(8\eta^{-1})\lambda^s\left(D\left(\lambda^r + \frac{1}{\sqrt{\ell'}}\right) + \left(\frac{R}{\ell'\lambda} + \sqrt{\frac{\Sigma^2\mathcal{N}(\lambda)}{\lambda\ell'}}\right)\right) \quad (2.12)$$

holds with $\ell' = \min\{\ell_2, \ell_3, \ell_4\}$, provided that $\ell' \geq \ell_0$ and all ℓ_i are as in Table 2.1.

We remark that the choice $s=0$ corresponds to the estimation error in the space $\mathcal{H}_k$, whereas $s=\frac{1}{2}$ corresponds to the prediction error in the space $L^2(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mu)$.

Finally, we establish asymptotic error bounds for the family of regularized estimators of the type (2.10), when learning from a stationary $\tau-$mixing sequence whose marginal distribution belongs to the class $\mathcal{M}(R, \Sigma, r, D, \beta, b)$, under an appropriate choice of the regularization parameter sequence λ_n. To simplify somewhat expressions, we will assume from now on, without loss of generality, that $D \geq R \geq 1$ holds. We separate the analysis between the cases of exponentially and polynomially $\tau-$mixing processes.

For an exponentially $\tau-$mixing process $(X_i, Y_i)_{i\geq 1}$ with mixing rate $\tau(k) = \chi\exp(-(\theta k)^{\gamma})$, we set:

$$\ell'_g(n) := \left\lfloor \frac{n\theta}{2(1 \vee \log(3nKD\chi\theta R^{-1}))^{\frac{1}{\gamma}}} \right\rfloor, \qquad \lambda_n := \min\left(\left(\frac{\Sigma^2}{D^2\ell'_g}\right)^{\frac{b}{2br+b+1}}, 1\right). \quad (2.13)$$

We observe in particular (by straightforward calculation, using the fact that $D \geq R \geq 1$) that the constraint $\ell'_g \leq \min\{\ell_2, \ell_3, \ell_4\}$ is fulfilled. We are then able to formulate the next statement.

Theorem 2.4.4. *Let distribution ν belong to the class $\mathcal{M}(R, \Sigma, r, D, \beta, b)$, and $f^{\lambda_n}_{\mathcal{D}_n}$ be a kernel spectral regularization estimator (2.10) with qualification $q \geq r+s$, where λ_n is given by (2.13). Fix some $\eta \in (0,1]$. Then there exists n_0 (depending on all the model parameters and on η) such that for $n \geq n_0$, it holds with probability at least $1-\eta$:*

$$\left\|T^s\left(f_\nu - f^{\lambda_n}_{\mathcal{D}_n}\right)\right\|_{\mathcal{H}_k} \leq C_*\log(8\eta^{-1})D\left(\frac{\Sigma}{D\sqrt{\ell'_g}}\right)^{\frac{2b(r+s)}{2br+b+1}}, \quad (2.14)$$

where $C_* := C_{r,s,B,E,\overline{\gamma},b,\beta}$ is a factor depending on the regularization function and model parameters (other than D, Σ).

We establish an analogous result for a polynomially $\tau-$mixing process $(X_i, Y_i)_{i\geq 1}$ with mixing rate $\tau(k) = \rho k^{-\gamma}$, this time without precisely tracking the effects of the constants (Σ, D). We also only consider the case of a two-sided controlled spectrum

$$\mathcal{P}^{\leqslant}(b, \beta_-, \beta_+) := \left\{\mu \in \mathcal{P} : \beta_- j^{-b} \leq \zeta_j \leq \beta_+ j^{-b}, \forall j \geq 1\right\},$$

and the model $\widetilde{\mathcal{M}}(R, \Sigma, r, D, \beta_\pm, b)$ defined as in (2.9) with $\mathcal{P}^<$ replaced by $\mathcal{P}^{\leqslant}$. The technical reason

for adding an assumption of lower bounded spectrum is that it implies a lower bound on the effective dimension $\mathcal{N}(\lambda)$, and in turn a lower bound on the effective sample size, which involves the effective dimension in the polynomial mixing case (see Table 2.1).

We consider the following parameter sequence:

$$\lambda_n := n^{-\frac{b}{2br+b+1+b(r+1)\gamma^{-1}}}. \tag{2.15}$$

Similarly to the case of exponential mixing, we use Lemma 2.4.3 with the choice $\ell'_p = \mathcal{O}\Big((\lambda_n \mathcal{N}(\lambda_n))^{\frac{2}{2\gamma+1}} n^{\frac{2\gamma}{2\gamma+1}}\Big)$, which depends on the regularization and on the effective dimension. Arguing in the same way as in Theorem 2.4.4 we then obtain:

Theorem 2.4.5. *Assume the data distribution ν belongs to the class $\widetilde{\mathcal{M}}(R, \Sigma, r, D, \beta_{\pm}, b)$, and $f_{\mathcal{D}_n}^{\lambda_n}$ is a kernel spectral regularization estimator (2.10) with qualification $q \geq r+s$, where λ_n is given by (2.15). For any fixed $\eta \in [0,1]$ and all $n > n_0$ (where n_0 is such that $\log(8\eta^{-1}) \leq C'_{\triangle} n_0^{\frac{br}{2br+b+1+b(r+1)\gamma^{-1}}}$, we have with probability at least $1-\eta$:*

$$\left\| T^s \left(f_\nu - f_{\mathcal{D}_n}^{\lambda_n} \right) \right\|_{\mathcal{H}_k} \leq C_{\triangle} \log(8\eta^{-1}) n^{-\frac{b(r+s)}{2br+b+1+b(r+1)\gamma^{-1}}}, \tag{2.16}$$

where $C_{\triangle}, C'_{\triangle}$ are factors depending on the regularization and smoothness parameters of the model $(R, \Sigma, D, r, s, B, E, \overline{\gamma}, b, \beta)$.

Let us briefly discuss the upper bounds for the risk of the general regularization methods, described in Theorems 2.4.4 and 2.4.5. Asymptotic in nature, these results are based on the concentration inequality (2.3.5), which allows the control of an error on the exponential scale. Comparing the result of Theorem 2.4.4 to risk bounds obtained for an i.i.d. scenario (e.g. in Blanchard and Mücke (2018)), we observe that in the case of an exponentially mixing process the upper bounds are optimal up to a logarithmic factor. In the case of a polynomially mixing process with exponent γ, the rate is degraded by a polynomial term that depends on γ. Naturally, it vanishes as $\gamma \to \infty$, as one would expect. In the case of exponential mixing, since we describe the explicit dependence of the sequence λ_n on Σ and D, further analysis can be conducted exploring other regimes in which either Σ or D may depend on n.

Remark 2.4.6. We consider a particular case of Sobolev RKHS $W_2^s(\mathcal{X})$ (see Chapter 4 for more details on Sobolev RKHS) in the case when $g_\nu \in W^s(\mathcal{X})$. In this case it is known that $b = \frac{2s}{d}$. To obtain the risk bounds in L_2-norm we take $s = \frac{1}{2}$ and since $f_\nu \in W^s(\mathcal{X})$ so $r = 0$. Rates from Theorem (2.4.4) for the geometrically decaying mixing coefficients imply that with probability at least $1-\eta$ any regularization method which outputs data-dependent decision rule $f_{\mathcal{D}_n,\lambda_n}$ learns regression function f_ν with rate (dropping the constant) $\left(\ell'_g\right)^{-\frac{s}{2s+d}}$. Since, (up to a logarithmic factor in n) $\ell'_b(n) = C_1 n$, where C_1 is some constant, so we have that the rates are essentially the same as the rates for the excess risk in the case of i.i.d. non-parametric regression over Sobolev spaces $W_2^s(\mathcal{X})$, $s > \frac{d}{2}$. It is known (see Blanchard and Mücke (2018) for the i.i.d. data observations) that the rates $n^{-\frac{s}{2s+d}}$ are optimal over the classes of Sobolev balls $W_2^s(\mathcal{X})$, thus we observe the logarithmic decay of the rate in the case of $\tau-$mixing dependency assumption.

2.4.3 Conclusions and perspectives

The results of Theorem 2.4.4 and 2.4.5 are stated in the somewhat standard framework of regularized learning schemes where the estimator takes values in a Hilbert space, which is generated by some reproducing kernel k.

Since the concentration results of Theorems 2.3.4 and 2.3.5 are valid in the more general case of complete normed spaces which satisfy the smoothness assumption **A1**, a natural extension is to consider the setting of statistical learning whose estimators are prediction functions belonging to a certain functional Banach space.

A variety of such Banach-valued learning schemes (using a corresponding Banach norm regularization) and theoretical justification of their validity have been proposed by numerous authors over the years. Seminal works Benett and Bredensteiner (2000) and Zhang (2002) have introduced extensions of convex risk regularization principles to involve a Banach norm regularizer terms. Further efforts have developed the mathematical foundations of such methods, in particular concerning properties of so-called evaluation Banach spaces or reproducing kernel Banach spaces and generalizations of the Representer theorem Canu et al. (2003), Hein et al. (2005) and Zhang et al. (2009) as well as universal approximation properties of such spaces, see Micchelli and Pontil (2004).

Furthermore, such approaches have given rise to numerous developments in recent research, for example extension to the vector-valued kernel setting (see Zhang and Zhang (2013)) with application for multi-task learning, the notion of orthomonotonicy, which leads to a generalization of representation theorems by Argyriou and Dinuzzo (2014), or combinations of these approaches with the kernel mean embedding principle (see Sriperumbudur et al. (2011)).

Concerning statistical properties (such as consistency, learning rates and generalization upper bounds for the risk) of Banach valued learning algorithms, these were also investigated in Hein et al. (2005), Steinwart (2009), Song and Zhang (2011) Combettes et al. (2018), albeit only for the case of independent training data.

Following Combettes et al. (2018),Zhang et al. (2009), as a direction for future work, one can investigate the geometrical properties of the underlying Banach space norm so as to ensure the possibility of learning in the normed space on the one hand, and to satisfy the smoothness assumption **A1** on the other. In such a situation the concentration results presented in this chapter will apply and have the potential to provide a major tool in the analysis of such schemes for weakly dependent data.

In Chapter 5 we investigate the deviation bounds (both in terms of exponential probability deviations and in L_p-norm, $p \geq 2$ norm) for the partial sums of weakly-dependent random fields indexed by the elements of grid in $\mathbb{N}^d$. Therein we obtain an improvement to the known results for the real-valued random fields (see Dedecker (1991)) which in particular case $d = 1$ implies known deviation bounds (both in L_p- norm and in probability on the exponential scale) obtained in the work Peligrad et al. (2007). We notice that techniques mentioned in Chapter 5 are based on the multidimensional multi scale martingale decomposition. They can be extended to the case of Banach-valued random sums and even in the case when the process is indexed by the elements in $\mathbb{N}$ probably lead to the improvement of the concentration bounds and thus to the algorithmic rates. One drawback is that the results of Chapter 5 are of Hoeffding (and not of Bernstein) type. This is however is due to the fact that we are using the analysis of sub-gaussian norm of the partial sum therein (which is equivalent to considering subgaussian deviations or, which is the same, inequalities of Hoeffding-type). The latter can be extended to the Bernstein's case by changing the subgaussian norm to a general type of Bernstein-Orlicz (or Benett-Orlicz) norm by doing the similar analysis and imposing additional weak-dependency assumption of L_2 type on the process $(X_t)_{t\in\mathbb{N}}$.

2.5 Proofs of the main probabilistic results

We will exploit the following auxiliary lemmata to give the proof of Theorem 2.3.5. We will use repeatedly the shorthand notation $\pi(x) := e^x - x - 1$.

Lemma 2.5.1. *Assume that $(X_i)_{i\geq 1}$ is a $\mathcal{C}-$mixing stochastic process with values in the closed subset $\mathcal{X} = B(c)$ of the separable Banach space $(\mathcal{B}, \|\cdot\|)$, such that Assumptions **A1**, **A2**, **A3** hold. Furthermore,*

let $(i_1, \dots, i_k)$ be a k-tuple of non-negative integers, such that $i_1 < i_2 \dots < i_k$, $\lambda \geq 0$ and $\tilde{S}_k := X_{i_1} + X_{i_2} + \dots + X_{i_k}$. Then the following holds:

$$\mathbb{E}\big[\exp\left(\lambda\|\tilde{S}_k\|\right)\big] \leq 2\left(1 + B\sigma^2\frac{\pi(\lambda c)}{c^2}\right)\prod_{j=1}^{k-1}(1 + p(d_j, \lambda))\,,$$

*where $p(k,\lambda) := \lambda\tilde{A}_1\phi_{\mathcal{C}}(k) + B\big(C_2\phi_{\mathcal{C}}(k) + \sigma^2\big)\frac{\pi(\lambda c)}{c^2}$, $d_j := i_j - i_{j-1}$ for all $j \geq 2$, $B := A_1^2 + A_2$, $\tilde{A}_1 = C_1A_1$ and constants A_1, A_2, C_1, C_2 as in assumptions **A1**, **A3**.*

Lemma 2.5.2. *Assume that $(X_i)_{i=1}^n$ is a random sample from a $\mathcal{X}$-valued centered $\phi_{\mathcal{C}}-$mixing process, such that Assumptions* **A1, A2, A3** *hold. For $n = \ell k + r$, where $\ell, k > 1$ are some integers and $r \in \{0, 1, \dots, k-1\}$, and any $\lambda \geq 0$, we have:*

$$\mathbb{E}\Big[\exp\Big(\lambda\Big\|\frac{1}{n}\sum_{i=1}^n X_i\Big\|\Big)\Big] \leq 2\exp\left(\frac{B}{c^2}\Big((\ell+1)\sigma^2 + C_2\ell\phi_{\mathcal{C}}(k)\Big)\pi\left(\frac{\lambda c}{\ell}\right) + \lambda\tilde{A}_1\phi_{\mathcal{C}}(k)\right). \tag{2.17}$$

where $\tilde{A}_1$ and B are defined as in Lemma 2.5.1.

Lemma 2.5.3. *If all the conditions of Lemma 2.5.2 hold then the following (exponential) inequality holds:*

$$\mathbb{P}\Big(\frac{1}{n}\Big\|\sum_{i=1}^n X_i\Big\| \geq t\Big) \leq 2\exp\left(-\frac{\ell\big(t^2 - 4\tilde{m}t\big)}{4\big(\frac{tc}{3} + \tilde{\sigma}^2B\big)}\right),$$

where $\tilde{m} := \tilde{A}_1\phi_{\mathcal{C}}(k)$ and $\tilde{\sigma}^2 := \sigma^2 + C_2\phi_{\mathcal{C}}(k)$.

Alternatively, for any $\nu > 0$ this can be written as:

$$\mathbb{P}\left[\Big\|\frac{1}{n}\sum_{i=1}^n X_i\Big\| \geq 4\tilde{m} + 4\sqrt{\frac{B\tilde{\sigma}^2\nu}{\ell}} + \frac{4}{3}\frac{c\nu}{\ell}\right] \leq 2\exp(-\nu).$$

Proof of Lemma 2.5.1

The backbone of the proof follows the technical approach as in the work Pinelis (1992). Use as a first step $\mathbb{E}[\exp(\lambda\|\tilde{S}_k\|)] \leq 2\mathbb{E}[\cosh(\lambda\|\tilde{S}_k\|)]$. The next step bounds iteratively the norm of $\tilde{S}_k$ by means of the norm of $\tilde{S}_{k-1}$ and additional terms which involve conditional expectation. To this end, we first need some (deterministic) bounds relating $\cosh(\lambda\|s+x\|)$ to $\cosh(\lambda\|s\|)$.

Let s, x be elements of $\mathcal{X}$. Introduce the following functions for $t \in [0,1]$:

$$f(t) := \cosh(\lambda h(t))\,, \qquad h(t) := \|s + tx\|\,.$$

For any $t \in [0,1]$ such that $h(t) \neq 0$, it holds

$$f'(t) = \lambda\sinh(\lambda h(t))h'(t) = \lambda\sinh(\lambda h(t))\big\langle D_{s+tx}\|\cdot\|, x\big\rangle\,. \tag{2.18}$$

If for some t_0, it holds $h(t_0) = 0$, then h itself may not be differentiable in t_0, however $f'(t)$ exists, and is equal to 0, in this case. Namely, if $x = 0$ then h must be identically zero and the claim follows. Otherwise $h(t) \neq 0$ for $t \neq t_0$, and Equation (2.18) holds for any $t \neq t_0$, implying by Assumption **A1** $|f'(t)| \leq A_1\lambda\|x\|\sinh(\lambda h(t))$; this implies differentiability in t_0 since the limit of the derivative exists (and is equal to 0) as $t \to t_0$, and the functions $h(t)$ and $f(t)$ are continuous. Similarly, for any $t \in [0,1]$

with $h(t) \neq 0$, and using Assumption **A1**:

$$\begin{aligned} f''(t) &= \lambda^2 \cosh(\lambda h(t)) h'(t)^2 + \lambda \sinh(\lambda h(t)) h''(t) \\ &= \lambda^2 \cosh(\lambda h(t)) \big\langle D_{s+tx} \|\cdot\|, x \big\rangle^2 + \lambda \sinh(\lambda h(t)) \delta_{x,x}\big(\big\|s+tx\big\|\big) \\ &\leq A_1^2 \lambda^2 \|x\|^2 \cosh(\lambda h(t)) + A_2 \lambda \|x\|^2 \frac{\sinh(\lambda h(t))}{h(t)} \\ &\leq \lambda^2 \|x\|^2 B \cosh(\lambda h(t)) \,, \end{aligned}$$

where we have used $\sinh(x) \leq x \cosh(x)$. We conclude that $f'(t)$ is absolutely continuous: unless $h(t)$ is identically 0, there exists at most a single point $t_0 \in [0,1]$ where $h(t_0) = 0$ and where f' may not be differentiable. We can therefore use the Taylor expansion:

$$f(1) = f(0) + f'(0) + \int_0^1 (1-t) f''(t) dt. \tag{2.19}$$

The integral rest can be bounded using previous inequality on f together with the triangle inequality, the elementary inequality $\cosh(a+b) \leq \cosh(a)\exp(b)$ for $b \geq 0$, and recalling that $\|x\| \leq c$:

$$\begin{aligned} \int_0^1 (1-t) f''(t) dt &\leq \lambda^2 \|x\|^2 B \int_0^1 (1-t) \cosh(\lambda(\|s\| + t\|x\|)) dt \\ &\leq \lambda^2 \|x\|^2 B \cosh(\lambda \|s\|) \int_0^1 (1-t) \exp(\lambda t c) dt \\ &= \|x\|^2 B \cosh(\lambda \|s\|) \frac{\pi(\lambda c)}{c^2} \,. \end{aligned}$$

Combining this with (2.19) and (2.18) we get for $s \neq 0$:

$$\cosh(\lambda\|s+x\|) = f(1) \leq \cosh(\lambda\|s\|)\Big(1 + \lambda \langle D_s \|.\|, x\rangle + \|x\|^2 B \frac{\pi(\lambda c)}{c^2}\Big) \,, \tag{2.20}$$

where we have used $\sinh a \leq \cosh a$ in (2.18). The above inequality remains true for $s = 0$ if we formally define $D_0\|.\|$ as 0, due to $f'(0) = 0$ in this case, as argued earlier.

We now go back to our initial goal of controlling $\mathbb{E}\big[\cosh\big(\lambda\big\|\tilde{S}_k\big\|\big)\big]$. We use the notation $\mathbb{E}_{j-1}[\cdot] := \mathbb{E}[\cdot|\mathcal{M}_{i_{j-1}}]$ where $M_{i_{j-1}} = \sigma(X_l : 1 \leq l \leq i_{j-1}), l \in \mathbb{N}$ using $s := \tilde{S}_{k-1}, x = X_{i_k}$ then taking conditional expectations in (2.20), we obtain

$$\begin{aligned} &\mathbb{E}_{k-1}\big[\cosh\big(\lambda\big\|\tilde{S}_k\big\|\big)\big] \\ &\qquad \leq \cosh\big(\lambda\big\|\tilde{S}_{k-1}\big\|\big)\left(1 + \lambda \mathbb{E}_{k-1}\big[\big\langle D_{\tilde{S}_{k-1}} \|.\|, X_{i_k}\big\rangle\big] + \mathbb{E}_{k-1}\big[\|X_{i_k}\|^2\big] B \frac{\pi(\lambda c)}{c^2}\right). \end{aligned} \tag{2.21}$$

In order to control the conditional expectation of the duality product on the right-hand side of (2.21), we will need the following measure-theoretical lemma:

Lemma 2.5.4. *Assume $\mathcal{X}, \mathcal{Y}, \mathcal{T}$ are three Polish spaces. Let F be a measurable real-valued function defined on $\mathcal{X} \times \mathcal{T}$, and let (X,Y) be a $\mathcal{X} \times \mathcal{Y}$-valued random variable (where $\mathcal{X} \times \mathcal{Y}$ is assumed to be equipped with the standard Borel sigma-algebra of open sets $\mathcal{B}(\mathcal{X} \times \mathcal{Y})$) on an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Denote through $B(t,\varepsilon)$ an open ball of radius ε, centered at point $t \in \mathcal{T}$. Assume that $F(X,t)$ is $\mathbb{P}$-integrable for all $t \in \mathcal{T}$ and that the following holds:*

- ***i)*** *For all $t \in \mathcal{T}$, $\|\mathbb{E}[F(X,t)|Y] - \mathbb{E}[F(X,t)]\|_\infty \leq C < \infty$;*
- ***ii)*** *The mapping $t \mapsto F(x,t)$ is continuous in t for all $x \in \mathcal{X}$;*

iii) *There exists $\varepsilon > 0$ and for all $t \in \mathcal{T}$ a measurable function $L_t(x) : \mathcal{X} \to \mathbb{R}_+$ such that for all $x \in \mathcal{X}$, $\sup_{t' \in B(t,\varepsilon)} |F(x,t')| \leq L_t(x)$, and $L_t(X)$ is $\mathbb{P}$-integrable.*

Then, there exist a version of the conditional expectations $\mathbb{E}[F(X,t)|Y]$ such that for $\mathbb{P}$-almost all y, we have:

$$\forall t \in \mathcal{T} \quad \big|\mathbb{E}[F(X,t)|Y=y] - \mathbb{E}\big[F(X,t)\big]\big| \leq C\,. \tag{2.22}$$

In particular, if $\mathcal{T} = \mathcal{Y}$, under the previous assumptions we conclude that

$$\Big\|\mathbb{E}[F(X,Y)|Y] - \mathbb{E}\big[F(\widetilde{X},Y)|Y\big]\Big\|_\infty \leq C\,, \tag{2.23}$$

where $\widetilde{X}$ is a copy of X which is independent of Y.

Observe that the whole point of this lemma is the inversion of quantificators "for all t, for almost all y" between its assumption (1) and the conclusion (2.22).

Proof of Lemma 2.5.4

Since $\mathcal{X}$ is Polish, there exists a regular conditional probability $\mathbb{P}(X \in \cdot|Y = \cdot)$, and we choose as a particular version of all conditional expectations the point-wise integral with respect to this stochastic kernel.

For every $t \in \mathcal{T}$ using the continuity of F in t, the dominated convergence theorem locally over a neighbourhood of a point t (Assumptions **ii)-iii)**) because $\mathcal{T}$ is Banach, we deduce that the function $t \mapsto \mathbb{E}[F(X,t)]$ is continuous over $\mathcal{T}$. Therefore, replacing F by $\widetilde{F}(x,t) := F(x,t) - \mathbb{E}[F(X,t)]$ and L_t by $2L_t$, we can assume without loss of generality that $\mathbb{E}[F(X,t)] = 0$ for all $t \in \mathcal{T}$. Since $\mathcal{T}$ is assumed to be Polish, in particular it is separable; let $\widetilde{\mathcal{T}}$ be a countable dense subset of $\mathcal{T}$. From assumption (1), for each $\tilde{t} \in \widetilde{\mathcal{T}}$ there exists a measurable set $A_{\tilde{t}} \subset \mathcal{Y}$ with $\mathbb{P}\big(Y \in A_{\tilde{t}}\big) = 1$, such that $\big|\int_{\mathcal{X}} F(x,\tilde{t})d\mathbb{P}(x|Y=y)\big| \leq C$ for all $y \in A_{\tilde{t}}$. Furthermore, for any $\tilde{t} \in \widetilde{\mathcal{T}}$, since the function $L_{\tilde{t}}(X)$ is $\mathbb{P}$-integrable, it holds $\int_{\mathcal{X}} L_{\tilde{t}}(x)d\mathbb{P}(x|Y=y) < \infty$ for all $y \in B_{\tilde{t}} \subset \mathcal{Y}$ with $\mathbb{P}\big(Y \in B_{\tilde{t}}\big) = 1$.

This together with countability implies that the set $A := \bigcap_{\tilde{t}\in\widetilde{\mathcal{T}}}(A_{\tilde{t}} \cap B_{\tilde{t}})$ is such that $\mathbb{P}\big(Y \in A\big) = 1$ and for all $(y,\tilde{t}) \in A \times \widetilde{\mathcal{T}}$, we have $\big|\int_{\mathcal{X}} F(x,\tilde{t})d\mathbb{P}(x|Y=y)\big| \leq C$ and $x \to L_{\tilde{t}}(x)$ is $\mathbb{P}(\cdot|Y=y)$-integrable.

For an arbitrary $t \in \mathcal{T}$, let $\tilde{t}_n$ be a sequence of points in $\widetilde{\mathcal{T}}$ converging to t in $\mathcal{T}$. We can assume without loss of generality that for all n, $d(\tilde{t}_n, t) < \varepsilon/2$ (where $\varepsilon > 0$ as in Assumption **iii)**), so that $d(\tilde{t}_n, \tilde{t}_{n'}) \leq \varepsilon$ for all n, n', implying that $\sup_n |F(x,t_n)| \leq L_t(x)$ holds (by Assumption **iii)**). Now for all $y \in A$, using continuity (Assumption 2) we have that for the version of conditional expectation under the regular conditional probability $\mathbb{P}(\cdot|Y=y)$ by dominated convergence it holds

$$\begin{aligned}\int_{\mathcal{X}} F(x,t)d\mathbb{P}(x|Y=y) &= \int_{\mathcal{X}} \lim_{n\to\infty} F(x,\tilde{t}_n)d\mathbb{P}(x|Y=y) \\ &= \lim_{n\to\infty} \int_{\mathcal{X}} F(x,\tilde{t}_n)d\mathbb{P}(x|Y=y) \leq C\,.\end{aligned}$$

In the case $\mathcal{T} = \mathcal{Y}$, we note that (2.22) implies (2.23) by choosing $t = y$. □

Returning now to the proof of Lemma 2.5.1, we use Lemma 2.5.4 with $\mathcal{Y} = \mathcal{X}^*$, $F(x,y) = \langle y, x\rangle$, and $(X,Y) = (X_{i_k}, D_{\tilde{S}_{k-1}}\|\cdot\|)$. By linearity of scalar product and expectation, and because the process $(X_i)_{i\geq 1}$ is centered, we have for fixed $y \in \mathcal{X}^*$: $\mathbb{E}[\langle y, X_{i_k}\rangle] = 0$. Obviously F is continuous in its first argument. Since by Assumption **A2**, $D_s\|\cdot\|$ is uniformly bounded and $\mathcal{X} = B(c)$, we can restrict the domain of F to $\mathcal{X} \times B^*(A_1)$, and F is then bounded uniformly, so that conditions (2) and (3) of Lemma 2.5.4 are satisfied. Because of Assumption **A3** it follows that $\|F(y,\cdot)\|_{\mathcal{C}} \leq C_1\|F(y,\cdot)\|_\infty \leq C_1A_1$. Finally, due to conditions on $\phi_{\mathcal{C}}$-mixing coefficients, we have that condition (1) is fulfilled with the constant $C = A_1C_1\phi_{\mathcal{C}}(d_k) := \tilde{A}_1\phi_{\mathcal{C}}(d_k)$, so from (2.23) we conclude, that:

$$\left|\mathbb{E}_{k-1}\left[\left\langle D_{\tilde{S}_{k-1}}\|.\|, X_{i_k}\right\rangle\right]\right| \leq \tilde{A}_1\phi(d_k)\,. \tag{2.24}$$

We turn to the control of the second conditional expectation on the right-hand side of (2.21). Using the $\phi_{\mathcal{C}}$-mixing assumption and Assumptions **A2, A3** again, we have almost surely (recalling $d_k := i_k - i_{k-1}$):

$$\begin{aligned}\mathbb{E}_{k-1}\left[\|X_{i_k}\|^2\right] &\leq \mathbb{E}_{k-1}\left[\|X_{i_k}\|^2\right] - \mathbb{E}\left[\|X_{i_k}\|^2\right] + \mathbb{E}\left[\|X_{i_k}\|^2\right] \\ &\leq C_2\phi_{\mathcal{C}}(d_k) + \sigma^2,\end{aligned}$$

and since by Assumption **A3**, the mapping $x \mapsto \|x\|^2$ is bounded in semi-norm $C(f)$ on $B(c)$ by some constant C_2 . Putting this bound together with (2.24) in the inequality (2.21), we get:

$$\mathbb{E}_{k-1}\left[\cosh\left(\lambda\left\|\tilde{S}_k\right\|\right)\right] \leq \cosh\left(\lambda\left\|\tilde{S}_{k-1}\right\|\right)(1 + p(d_k, \lambda))\,,$$

where we recall $p(k,\lambda) := \lambda\tilde{A}_1\phi_{\mathcal{C}}(k) + B\big(C_2\phi_{\mathcal{C}}(k) + \sigma^2\big)\left(\frac{\pi(\lambda c)}{c^2}\right)$.

Iteratively repeating the aforementioned argument and considering that the bound on conditional expectation $\mathbb{E}_{k-1}[\cdot]$ holds almost surely, one obtains:

$$\begin{aligned}\mathbb{E}\left[\cosh\left(\lambda\left\|\tilde{S}_k\right\|\right)\right] &= \mathbb{E}\left[\mathbb{E}_{k-1}\left[\cosh\left(\lambda\left\|\tilde{S}_k\right\|\right)\right]\right] \\ &\leq \mathbb{E}\left[\cosh\left(\lambda\left\|\tilde{S}_{k-1}\right\|\right)\right](1 + p(d_k, \lambda)) \\ &\leq \mathbb{E}\left[\cosh\left(\lambda\left\|X_{i_1}\right\|\right)\right]\prod_{j=2}^{k}(1 + p(d_j, \lambda)).\end{aligned}$$

For bounding $\mathbb{E}\left[\cosh\left(\lambda\left\|X_{i_1}\right\|\right)\right]$ we use (2.20) with $s = 0$ and obtain:

$$\mathbb{E}[\cosh\|X_{i_1}\|] \leq \mathbb{E}\left[1 + \|X_{i_1}\|^2 B\frac{\pi(\lambda c)}{c^2}\right] \leq 1 + \frac{\sigma^2}{c^2}B\pi(\lambda c)\,,$$

which implies the claim. □

To proceed in the proof, we use the classical (see for example in Bosq (1993),Wintenberger (2010) and Hang and Steinwart (2017)) approach and divide the sample $(X_1, \ldots, X_n)$ into blocks, such that the distance between two neighbor elements in a given block will be large enough to ensure small dependence coefficient. We partition the set $\{1, 2, \ldots, n\}$ into k blocks in the following way. Write $n = \ell k + r, 0 \leq r \leq k - 1$ and define

$$I_i = \begin{cases} \{i, i + k, \ldots, i + \ell k\}, & \text{if } 1 \leq i \leq r\,, \\ \{i, i + k, \ldots, i + (\ell - 1)k\}, & \text{if } r + 1 \leq i \leq k\,. \end{cases}$$

Denote through $|I_i|$ the number of elements in the $i-$th block; it holds $|I_i| = \ell + 1$ for $1 \leq i \leq r$, $|I_i| = \ell$ for $r + 1 \leq i \leq k$, and $\sum_{i=1}^{k}|I_i| = n$. We use the notation $S_{I_i} = \sum_{j\in I_i} X_j$. Now we use Lemma 2.5.1 for each of the constructed blocks I_i, $1 \leq i \leq k$ to prove Lemma 2.5.2.

Proof of Lemma 2.5.2

By the triangle inequality $\|S_n\| \leq \sum_{j=1}^{k}\left\|S_{I_j}\right\|$, implying for any $\lambda > 0$, via the convexity of the exponential function:

$$\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\|S_n\|\right)\right] \leq \mathbb{E}\left[\exp\left(\lambda\sum_{j=1}^{k} r_j\frac{\left\|S_{I_j}\right\|}{|I_j|}\right)\right] \leq \sum_{j=1}^{k} r_j\mathbb{E}\left[\exp\left(\frac{\lambda}{|I_j|}\left\|S_{I_j}\right\|\right)\right], \tag{2.25}$$

where $r_j := \frac{|I_j|}{n}$, with $\sum_{j=1}^{k} r_j = 1$. Now for each summand in the last sum, we apply Lemma 2.5.1 for

the index tuple given by the ordered elements of I_j, yielding

$$\mathbb{E}\Big[\exp\Big(\frac{\lambda}{|I_j|}\|S_{I_j}\|\Big)\Big] \le 2\Big(1+B\frac{\sigma^2}{c^2}\pi\Big(\frac{\lambda c}{|I_j|}\Big)\Big)\Big(1+p\Big(k,\tfrac{\lambda}{|I_j|}\Big)\Big)^{|I_j|-1}.$$

Substituting the last bound into (2.25), we obtain:

$$\begin{aligned}\mathbb{E}\Big[\exp\Big(\frac{\lambda}{n}\|S_n\|\Big)\Big] &\le 2\sum_{j=1}^{k} r_j\Big(1+B\frac{\sigma^2}{c^2}\pi\Big(\frac{\lambda c}{|I_j|}\Big)\Big)\Big(1+p\Big(k,\tfrac{\lambda}{|I_j|}\Big)\Big)^{|I_j|-1}\\ &\le 2\sum_{j=1}^{k} r_j \exp\Big(\frac{B\sigma^2}{c^2}\pi\Big(\frac{\lambda c}{\ell}\Big)\Big)\exp\big(\ell p\big(k,\tfrac{\lambda}{\ell}\big)\big),\end{aligned}$$

where we used the simple bound $1+x \le \exp(x)$ twice, the condition $\ell \le |I_j| \le \ell+1$, and the fact that $p(k,\cdot)$ is non-decreasing in function for fixed k. The last quantity is equivalent to the claim of the lemma. □

Proof of Lemma 2.5.3

Using Chernoff's bound and Lemma 2.5.2, we obtain for any $\lambda > 0$:

$$\begin{aligned}\mathbb{P}\Big[\tfrac{1}{n}\|S_n\| \ge t\Big] &= \mathbb{P}\Big[\exp\big(\tfrac{1}{n}\|\lambda S_n\|\big) \ge \exp(\lambda t)\Big]\\ &\le \exp(-\lambda t)\mathbb{E}\Big[\exp\big(\tfrac{\lambda}{n}\|S_n\|\big)\Big]\\ &\le 2\exp\Big(-\lambda(t-\tilde{m}) + \tilde{\sigma}^2\frac{(\ell+1)B}{c^2}\pi\Big(\frac{\lambda c}{\ell}\Big)\Big),\end{aligned} \tag{2.26}$$

where $\tilde{m} := \tilde{A}_1\phi_{\mathcal{C}}(k)$ and $\tilde{\sigma}^2 := \sigma^2 + C_2\phi_{\mathcal{C}}(k)$.

First we get an upper bound on the value of the function $\pi(\frac{\lambda c}{l})$. By using the Taylor series decomposition, simple inequality $2\cdot 3^{k-2} \le k!$ for $k \in \mathbb{N}$ and summing the geometric series we obtain:

$$\pi\Big(\frac{\lambda c}{\ell}\Big) \le \sum_{j=2}^{\infty}\Big(\frac{\lambda c}{\ell}\Big)^j\frac{1}{2\cdot 3^{j-2}} = \frac{\lambda^2c^2}{2\ell^2}\frac{1}{1-\frac{\lambda c}{3\ell}},$$

where we assume that $0 < \lambda < \frac{3\ell}{c}$. Inserting this inequality into (2.26) and simplifying the terms we get:

$$\mathbb{P}\Big[\tfrac{1}{n}\|S_n\| \ge t\Big] \le 2\exp\Big(-\lambda(t-\tilde{m}) + \tilde{\sigma}^2\lambda^2\frac{3(\ell+1)B}{2\ell}\frac{1}{3\ell-\lambda c}\Big). \tag{2.27}$$

Now we put $\lambda = \frac{t\ell}{\frac{tc}{3}+\tilde{\sigma}^2B}$. Clearly, by this choice of λ we have:

$$\frac{\lambda}{\ell} = \frac{t}{\frac{tc}{3}+\tilde{\sigma}^2B} \le \frac{3}{c}.$$

Thus, the choice of λ satisfies the assumption; putting it into the exponent of the right hand side of (2.27)

we then obtain:

$$
\begin{aligned}
-\lambda(t-\tilde{m})+\frac{3}{2}\tilde{\sigma}^2\frac{(\ell+1)B}{\ell}\lambda^2\frac{1}{3\ell-\lambda c} & \\
&= -\frac{t\ell(t-m)}{\frac{tc}{3}+\tilde{\sigma}^2B}+\frac{3}{2}\tilde{\sigma}^2\frac{(\ell+1)B}{\ell}\frac{t^2\ell^2}{\left(\frac{tc}{3}+\tilde{\sigma}^2B\right)^2}\frac{1}{3\ell-\frac{t\ell c}{\frac{tc}{3}+\tilde{\sigma}^2B}} \\
&= -\frac{t\ell(t-\tilde{m})}{\frac{tc}{3}+\tilde{\sigma}^2B}+\frac{1}{2}\frac{(\ell+1)t^2}{\frac{tc}{3}+\tilde{\sigma}^2B} \\
&= -\frac{(\ell-1)t^2-2\ell\tilde{m}t}{2\left(\frac{tc}{3}+\tilde{\sigma}^2B\right)}.
\end{aligned}
$$

Putting this into the exponent bound and upper bounding ℓ with $2(\ell-1)$, for $\ell \geq 2$, we get the claim of the lemma. □

Proof of Theorem 2.3.4

From the very last claim of Lemma 2.5.3 we have:

$$
\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}X_i\right\| \geq t\right] \leq 2\exp\left(-\frac{\ell\left(t^2-4\tilde{m}t\right)}{4\left(\frac{tc}{3}+\tilde{\sigma}^2B\right)}\right).
$$

Setting $\frac{\ell\left(t^2-4\tilde{m}t\right)}{4\left(\frac{tc}{3}+\tilde{\sigma}^2B\right)} := \nu$ and solving the last equation in terms of t, we obtain:

$$
\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}X_i\right\| \geq 4\tilde{A}_1\phi_{\mathcal{C}}(k)+4\sqrt{\frac{B\tilde{\sigma}^2\nu}{\ell}}+\frac{4}{3}\frac{c\nu}{\ell}\right] \leq 2\exp(-\nu), \tag{2.28}
$$

which proves the claim of the theorem. □

Proof of Theorem 2.3.5

From Theorem 2.3.4, assuming the *effective sample size* $\ell^* \geq 2$ defined as in Equation (2.5), putting $C_* = C_2/C_1$, we obtain that with probability at least $1-2\exp(-\nu)$ it holds:

$$
\left\|\frac{1}{n}\sum_{i=1}^{n}X_i\right\| \leq 4A_1\left(\frac{c}{\ell^*}\vee\frac{\sigma}{\sqrt{\ell^*}}\right)+4\sqrt{\frac{B\left(\sigma^2+C_*\left(\frac{c}{\ell^*}\vee\frac{\sigma}{\sqrt{\ell^*}}\right)\right)\nu}{\ell^*}}+\frac{4}{3}\frac{c\nu}{\ell^*} =: \tilde{L}. \tag{2.29}
$$

For $a, b > 0$ using the obvious inequalities $a \vee b \leq a+b$ and $\sqrt{ab} \leq (a+b)/2$ we obtain:

$$
\begin{aligned}
\tilde{L} &\leq 4A_1\left(\frac{c}{\ell^*}+\frac{\sigma}{\sqrt{\ell^*}}\right)+4\frac{\sqrt{B\nu}\sigma}{\sqrt{\ell^*}}+4\frac{\sqrt{BC_*c\nu}}{\ell^*}+4\frac{\sqrt{BC_*\sigma\nu}}{\sqrt{\ell^*}\sqrt{\ell^*}}+\frac{4}{3}\frac{c\nu}{\ell^*} \\
&\leq 4A_1\left(\frac{c}{\ell^*}+\frac{\sigma}{\sqrt{\ell^*}}\right)+4\frac{\sqrt{B\nu}\sigma}{\sqrt{\ell^*}}+2\frac{\sqrt{B\nu}(C_*+c)}{\ell^*}+2\sqrt{B\nu}\left(\frac{C_*}{\ell^*}+\frac{\sigma}{\sqrt{\ell^*}}\right)+\frac{2c\nu}{\ell^*} \\
&\leq \frac{\sigma}{\sqrt{\ell^*}}\left(4A_1+6\sqrt{B\nu}\right)+\frac{c}{\ell^*}\left(2\nu+2\sqrt{B\nu}+4A_1+4\sqrt{B\nu}\frac{C_*}{c}\right).
\end{aligned}
$$

Finally, we observe that the inequality

$$
\left\|\frac{1}{n}\sum_{i=1}^{n}X_i\right\| \leq \frac{\sigma}{\sqrt{\ell^*}}\left(4A_1+6\sqrt{B\nu}\right)+\frac{c}{\ell^*}\left(2\nu+2\sqrt{B\nu}+4A_1+4\sqrt{B\nu}\frac{C_*}{c}\right),
$$

trivially holds also for $\ell^* = 1$, since $A_1 \geq 1$. This implies the statement of the theorem using $\sqrt{\nu} \leq \nu$,

since we assumed $\nu \geq 1$ here.

□

We are now have all necessary technical tools in order to prove the exponential bounds for different decay rates of the mixing coefficients.

Proof of Proposition 2.3.6 We choose a reasonable bound ℓ_g on the effective sample size ℓ^* in the case of geometrical mixing. Since $\phi_{\mathcal{C}}(\cdot)$ (extended to the positive real line as $\phi_{\mathcal{C}}(t) = \chi \exp(-(\theta t)^\gamma)$) is nonincreasing and $\frac{n}{2\ell} \leq \lfloor \frac{n}{\ell} \rfloor$, it is sufficient to choose ℓ_g such that $C_1 \phi_{\mathcal{C}}\left(\frac{n}{2\ell_g}\right)$ is smaller than $\frac{c}{\ell_g} \vee \frac{\sigma}{\sqrt{\ell_g}}$. Moreover, in the case of geometrical mixing, it is sufficient to choose ℓ_g such that $C_1 \phi_{\mathcal{C}}\left(\frac{n}{2\ell_g}\right) < \frac{c}{\ell_g}$ (trivially this implies that $C_1 \phi_{\mathcal{C}}\left(\frac{n}{2\ell_g}\right) < \frac{c}{\ell_g} \vee \frac{\sigma}{\sqrt{\ell_g}}$). We choose $\ell_g = \left\lfloor \frac{n\theta}{2(1 \vee \log(n\theta\chi C_1/c))^{1/\gamma}} \right\rfloor$. It is easy to check that in this case, we get

$$\ell_g C_1 \phi_{\mathcal{C}}\left(\frac{n}{2\ell_g}\right) \leq n\theta\chi C_1 \exp\left(-\left(1 \vee \log \frac{\chi\theta C_1 n}{c}\right)\right) \leq c,$$

which together with the result of Theorem 2.3.5 implies the first claim of the proposition.

For the case of polynomially mixing process, we have the coefficient decay rate $\phi_{\mathcal{C}}(k) = \rho k^{-\gamma}$. Similarly, we choose a bound ℓ_p for the *effective sample size* ℓ^* so that the conditions of Theorem 2.3.5 are satisfied. Analogously, it is sufficient to choose ℓ_p such that $C_1 \phi_{\mathcal{C}}\left(\frac{n}{2\ell_p}\right) \leq \frac{\sigma}{\sqrt{\ell_p}} \vee \frac{c}{\ell_p}$. Solving $C_1 \phi_{\mathcal{C}}\left(\frac{n}{2\ell}\right) \leq \frac{\sigma}{\sqrt{\ell}} \vee \frac{c}{\ell}$ in ℓ for given n, σ, c, ρ, C_1 results in the following choice:

$$\ell_p = \max\left\{ \left\lfloor \left(\frac{\sigma}{C_1\rho}\right)^{\frac{2}{2\gamma+1}} \left(\frac{n}{2}\right)^{\frac{2\gamma}{2\gamma+1}} \right\rfloor, \left\lfloor \left(\frac{c}{C_1\rho}\right)^{\frac{1}{\gamma+1}} \left(\frac{n}{2}\right)^{\frac{\gamma}{\gamma+1}} \right\rfloor \right\}, \tag{2.30}$$

which matches the claim of the Proposition. □

2.6 Proofs of the main statistical learning results.

As mentioned in section 2.4.2, in order to make use of the concentration inequalities for sums of random variables in $\mathrm{HS}(\mathcal{H}_k)$ and in $\mathcal{H}_k$, we should ensure that the functions of interest of the original $\tau-$mixing process $Z_i = (X_i, Y_i)$ are again $\tau-$mixing. This claim is established by the proof of the Lipschitz property of the corresponding mappings in the next lemma.

Lemma 2.6.1. *Assume $\mathcal{H}_k$ is a RKHS over $\mathcal{X}$, which is assumed to be a closed subset of a Polish space. Let the reproducing kernel k of $\mathcal{H}_k$ satisfies $\sup_{x\in\mathcal{X}} \sqrt{k(x,x)} \leq 1$. Assume that the the kernel admits a mixed partial derivative (in Gâteaux sense), $\partial_{1,2}k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ which is uniformly bounded on $\mathcal{X}$ by some constant $K > 0$. Finally, let $\mathcal{Y} = [-R, R]$. Then, the mapping $V : \mathcal{X} \to \mathrm{HS}(\mathcal{H}_k) : x \mapsto k_x \otimes k_x^*$ is $2K$-Lipschitz; for a fixed $f \in \mathcal{H}_k$, the mapping $W_f : \mathcal{X} \times \mathcal{Y} \to \mathcal{H}_k : (x, y) \mapsto yk_x - k_x\langle k_x, f\rangle$ is $3\max(KR, K\|f\|, 1)$-Lipschitz.*

Proof of Lemma 2.6.1 As a starting point, because of the assumption of uniform boundedness of the (mixed) partial derivative of the kernel k and Lemma 3.3 from Blanchard et al. (2011), we deduce that

k_x is $K-$Lipschitz as a map $\mathcal{X} \to \mathcal{H}_k$. Then, for arbitrary x_1, x_2 we obtain:

$$\begin{aligned}
\left\|k_{x_1} \otimes k_{x_1}^* - k_{x_2} \otimes k_{x_2}^*\right\|_{\mathrm{HS}(\mathcal{H}_k)}^2 &= \left\|k_{x_1} \otimes k_{x_1}^* - k_{x_1} \otimes k_{x_2}^* + k_{x_1} \otimes k_{x_2}^* - k_{x_2} \otimes k_{x_2}^*\right\|_{\mathrm{HS}}^2 \\
&= \left\|k_{x_1} \otimes \left(k_{x_1}^* - k_{x_2}^*\right)\right\|_{\mathrm{HS}}^2 + \left\|(k_{x_1} - k_{x_2}) \otimes k_{x_2}^*\right\|_{\mathrm{HS}}^2 \\
&\quad + 2\left\langle k_{x_1} \otimes \left(k_{x_1}^* - k_{x_2}^*\right), (k_{x_1} - k_{x_2}) \otimes k_{x_2^*}\right\rangle \\
&\leq \|k_{x_1}\|_{\mathcal{H}_k}^2 \left\|k_{x_1}^* - k_{x_2}^*\right\|_{\mathcal{H}_k}^2 + \|k_{x_1} - k_{x_2}\|_{\mathcal{H}_k}^2 \left\|k_{x_2}^*\right\|_{\mathcal{H}_k}^2 \\
&\quad + 2\sqrt{\|k_{x_1}\|_{\mathcal{H}_k}^2 \|k_{x_2}\|_{\mathcal{H}_k}^2} \|k_{x_1} - k_{x_2}\|_{\mathcal{H}_k}^2 \\
&\leq 4K^2 \|x_1 - x_2\|^2,
\end{aligned} \tag{2.31}$$

where we used the properties of the Hilbert-Schmidt norm of tensor product operators, the Cauchy-Schwartz inequality in the third line, the assumptions about boundedness of the kernel $\|k_x\|_{\mathcal{H}_k}^2 = k(x,x) \leq 1$ and that fact the map $x \mapsto k_x$ is $K-$Lipschitz in the last line.

Thus the map $V(x) = k_x\langle k_x, \cdot\rangle$ is $2K$-Lipschitz. Furthermore, we deduce

$$\|k_{x_1}\langle k_{x_1}, f\rangle - k_{x_2}\langle k_{x_2}, f\rangle\| \leq \left\|k_{x_1} \otimes k_{x_1}^* - k_{x_2} \otimes k_{x_2}^*\right\|_{\mathrm{HS}} \|f\| \leq 2K\|f\|\|x_1 - x_2\|. \tag{2.32}$$

Quite analogously, for any $(x_1, y_1), (x_2, y_2) \in \mathcal{X} \times \mathcal{Y}$ we have:

$$\begin{aligned}
\|y_1 k_{x_1} - y_2 k_{x_2}\|_{\mathcal{H}_k} &= \|y_1 k_{x_1} - y_1 k_{x_2} + y_1 k_{x_2} - y_2 k_{x_2}\|_{\mathcal{H}_k} \\
&= \|y_1(k_{x_1} - k_{x_2}) + k_{x_2}(y_1 - y_2)\|_{\mathcal{H}_k} \\
&\leq KR\|x_1 - x_2\|_{\mathcal{X}} + |y_1 - y_2| \\
&\leq \max(KR, 1)(\|x_1 - x_2\|_{\mathcal{X}} + |y_1 - y_2|).
\end{aligned} \tag{2.33}$$

The latter implies that the map $(x, y) \mapsto yk_x$ is $\max(KR, 1)$-Lipschitz. By gathering bounds from (2.32) and (2.33), we deduce that the random variables $W_f(x, y) := yk_x - k_x\langle k_x, f\rangle$ are Lipschitz with constant $3\max(1, KR, K\|f\|)$ as a map $\mathcal{X} \times \mathcal{Y} \to \mathcal{H}_k$. □

Proof of Lemma 2.4.1

Consider the mapping

$$\xi_1(x, y) := yk_x - k_x\langle k_x, f_\nu\rangle,$$

with values in $\mathcal{H}_k$. It holds $\frac{1}{n}\sum_{i=1}^n \xi(x_i, y_i) = S_n^* y - T_{\mathbf{x}} f_\nu$, as well as

$$\mathbb{E}[\xi_1(X, Y)] = \mathbb{E}[k_x(y - \langle f, k_x\rangle)] = \int_{\mathcal{X}} k_x \int_{\mathcal{Y}} (y - f_\nu(x))\nu(dy|x)\mu(dx) = 0$$

.

By Cauchy-Schwarz, reproducing property and Assumption **B1** we have

$$\|\xi_1(x, y)\| = \|yk_x - k_x\langle k_x, f_\nu\rangle\|_{\mathcal{H}_k} \leq \|k_x\||y - f_\nu(x)| \leq 2R.$$

Similar, due to Assumption **B1** and since $\sup_{x\in\mathcal{X}} k(x, x) \leq 1$, we obtain the following bound on the variance:

$$\begin{aligned}
\mathbb{E}\left[\|\xi_1(X, Y)\|^2\right] &= \int_{\mathcal{X}\times\mathcal{Y}} \langle k_x(y - f_\nu(x)), k_x(y - f_\nu(x))\rangle d\nu(x, y) \\
&= \int_{\mathcal{X}} d\mu(x) k(x, x) \int_Y (y - f_\nu(x))^2 d\nu(y|x) \leq \Sigma^2.
\end{aligned}$$

By Lemma 2.6.1 applied to $\xi_1(x,y) = W_{f_\nu}(x,y)$ we have that if $(X_i,Y_i)_{i\geq 1}$ is $\tau-$mixing with rate $\tau(k)$, the sequence $\xi_1(x_i,y_i)_{i\geq 1}$is $\tau-$mixing with rate $\tau(k) = 3\max(1,KR,KD)\tau(k)$. Using the result of Corollary 2.3.7 with the aforementioned bounds on the norm, the variance and the multiplicative correction for the mixing coefficients decay rate, we obtain with probability at least $1-\eta$ it holds:

$$\|S_k{}^* y - Tf_\nu\| \leq 21\log(2\eta^{-1})\left(\frac{\Sigma}{\sqrt{\ell_1}} + \frac{2R}{\ell_1}\right),$$

where the bound on the *effective sample size* ℓ_1 is obtained by a direct plug-in of the bounds for the norm, the second moment and the form of mixing coefficients of the sequence $\xi_1(x_i,y_i)$ in the general form given by Proposition 2.3.6. Namely, we have $\ell_1 = \left(\frac{\Sigma}{3\max(1,KR,KD)))\rho}\right)^{\frac{2}{2\gamma+1}}\left(\frac{n}{2}\right)^{\frac{2\gamma}{2\gamma+1}}$, for a polynomially mixing process with rate $\tau(k) = \rho k^{-\gamma}$, and $\ell_1 = \left\lfloor \frac{n\theta}{2\left(1\vee\log\left(n\frac{3\max(1,KR,KD)\chi\theta}{2R}\right)\right)^{\frac{1}{\gamma}}}\right\rfloor$ for an exponentially mixing process with rate $\tau(k) = \chi\exp(-(\theta k)^\gamma)$. The other inequalities will be derived in a similar way. We introduce the random variable:

$$\xi_2(x,y) = (T+\lambda)^{-\frac{1}{2}}(k_x y - k_x\langle k_x, f_\nu\rangle),$$

Quite analogously, we can check that $\mathbb{E}[\xi_2(X,Y)] = 0$. Repeating similar steps as in the first case, we get:

$$\left\|(T+\lambda)^{-\frac{1}{2}}(k_x y - k_x\langle k_x, f_\nu\rangle)\right\|_{\mathcal{H}_k} \leq \left\|(T+\lambda)^{-\frac{1}{2}}\right\|\|(k_x y - k_x\langle k_x, f_\nu\rangle)\|_{\mathcal{H}_k} \leq 2\lambda^{-\frac{1}{2}}R.$$

For the second moment of the norm of $\xi_2(X,Y)$, we get:

$$\begin{aligned}
\mathbb{E}\left[\|\xi_2(X,Y)\|^2\right] &= \int_{\mathcal{X}\times\mathcal{Y}}\left\langle (T+\lambda)^{-\frac{1}{2}}k_x(y-f_\nu(x)), (T+\lambda)^{-\frac{1}{2}}k_x(y-f_{\mathcal{H}_k}(x))\right\rangle d\nu(x,y)\\
&= \int_{\mathcal{X}}\left\|(T+\lambda)^{-\frac{1}{2}}k_x\right\|^2 d\mu(x)\int_{\mathcal{Y}}(y-f_\nu(x))^2 d\nu(y|x)\\
&\leq \Sigma^2\int_{\mathcal{X}} Tr\left((T+\lambda)^{-\frac{1}{2}}k_x\otimes k_x^*\right)d\mu(x)\\
&= \left(\Sigma\sqrt{\mathcal{N}(\lambda)}\right)^2.
\end{aligned}$$

By Lemma 2.6.1, one has that the function $\xi_2(x,y) = (T+\lambda)^{-\frac{1}{2}}W_{f_\nu}(x,y)$ is Lipschitz with constant $3\lambda^{-\frac{1}{2}}\max(1,KR,KD)$, from which we deduce that $(\xi_2(X_i,Y_i))_{i\geq 1}$ is $\tau-$mixing with rate

$$3\lambda^{-\frac{1}{2}}\max(1,KR,KD)\tau(k)$$

. Finally, by using Corollary 2.3.7, we obtain with probability at least $1-\eta$:

$$\left\|(T+\lambda)^{-\frac{1}{2}}(T_n f_{\mathcal{H}_k} - S_n^* y)\right\|_{\mathcal{H}_k} \leq 21\log\left(\frac{2}{\eta}\right)\kappa^{-1}\left(\frac{\Sigma\sqrt{\mathcal{N}(\lambda)}}{\sqrt{\ell_2}} + \frac{2R}{\sqrt{\lambda}\ell_2}\right),$$

where, as before, a bound on ℓ_2 is obtained by Proposition 2.3.6 for either a polynomially or exponentially mixing process, through considering bounds on the norm, the second moment and the Lipschitz norm of the elements of the sequence $\xi_2(x_i,y_i)$.

We define the map $\xi_3 : \mathcal{X}\mapsto \mathrm{HS}(\mathcal{H})$ (here, as mentioned before, through $\mathrm{HS}(\mathcal{H})$ we denote the

space of Hilbert-Schmidt operators on $\mathcal{H}_k$) by:

$$\xi_3(x) := (T+\lambda)^{-1}(T_n - T),$$

where we recall the notation $T_n := k_x \otimes k_x^*$ for any $x \in \mathcal{X}$. Taking the expectation we get:

$$\mathbb{E}[\xi_3(X)] = (T+\lambda)^{-1}\int_{\mathcal{X}}(T_n - T)d\mu(x) = 0.$$

So that we have:

$$\Big\|(T+\lambda)^{-1}(T-T_n)\Big\|_{\mathrm{HS}(\mathcal{H}_k)} = \left\|\frac{1}{n}\sum_{i=1}^{n}\xi_3(x_i)\right\|_{HS(\mathcal{H}_k)}.$$

Verifying the conditions for the uniform bound and variance as above we obtain:

$$\|\xi_3(x)\|_{\mathrm{HS}} \leq \Big\|(T+\lambda)^{-1}\Big\|\|T-T_n\|_{\mathrm{HS}(\mathcal{H}_k)} \leq 2\lambda^{-1},$$

where $\Big\|(T+\lambda)^{-1}\Big\|$ is the supremum norm of the operator $(T+\lambda)^{-1}$. For the variance we have

$$\begin{aligned}
\mathbb{E}\Big[\|\xi_3(X)\|_{\mathrm{HS}}^2\Big] &= \int_{\mathcal{X}} Tr\Big((T_n - T)(T+\lambda)^{-2}(T_n - T)\Big)\mu(dx) \\
&= \int_{\mathcal{X}} Tr\Big(T_n(T+\lambda)^{-2}T_n\Big)\mu(dx) - Tr\Big(T(T+\lambda)^{-2}T\Big) \\
&\leq \|T+\lambda\|^{-1}\int_{\mathcal{X}}\|T_n\|Tr\Big((T+\lambda)^{-1}T_n\Big)\mu(dx) \\
&\leq \lambda^{-1}\mathcal{N}(\lambda).
\end{aligned}$$

By Lemma 2.6.1 ξ_3 is Lipschitz with constant $2\lambda^{-1}K$, thus $(\xi_3(X_i, Y_i))_{i\geq 1}$ is $\tau-$mixing with rate $2\lambda^{-1}K\tau(k)$.

We apply Theorem 2.3.5 to the quantity $\big\|\frac{1}{n}\sum_{i=1}^{n}\xi_3(x_i)\big\|_{\mathrm{HS}(\mathcal{H}_k)}$. With probability at least $1-\eta$ we have:

$$\Big\|(T+\lambda)^{-1}(T-T_n)\Big\|_{\mathrm{HS}(\mathcal{H}_k)} \leq 21\log\left(\frac{2}{\eta}\right)\left(\frac{\lambda^{-1/2}\sqrt{\mathcal{N}(\lambda)}}{\sqrt{\ell_3}} + \frac{2\lambda^{-1}}{\ell_3}\right).$$

where ℓ_3 is chosen following the standard "plug-in" scheme as before.

Finally, define $\xi_4(x) := (k_x \otimes k_x^* - T)$. Again the random variables $\xi_4(X_i)$ are centered and we have:

$$T_n - T = \frac{1}{n}\sum_{i=1}^{n}\xi_4(x_i).$$

Repeating the scheme we get:

$$\begin{aligned}
\|\xi_4(X)\|_{\mathrm{HS}(\mathcal{H})} &\leq 2, \\
\mathbb{E}\Big[\|\xi_4(X)\|_{\mathrm{HS}(\mathcal{H})}^2\Big] &\leq 4,
\end{aligned}$$

Also, Lemma 2.6.1 implies that $\xi_4(X_i)_{i\geq 0}$ is $\tau-$mixing with rate $2K\tau(k)$, so that using the general deviation bound from Corollary 2.3.7 and according to the same principle as above, we obtain with

probability at least $1-\eta$:

$$\|T - T_n\|_{\mathrm{HS}(\mathcal{H}_k)} \leq 21 \log\left(\frac{2}{\eta}\right)\left(\frac{2}{\sqrt{\ell_4}} + \frac{2}{\ell_4}\right) \leq \frac{42\log\left(2\eta^{-1}\right)}{\sqrt{\ell_4}},$$

where ℓ_4 is chosen according to the mixing rate and bounds on the norm, variance term and Lipschitz constants as above. □

Next we give an auxiliary lemma, in the same spirit as the i.i.d. counterparts in Blanchard and Mücke (2018).

Lemma 2.6.2. *Assume the conditions of Lemma 2.4.1 are satisfied. Let $\eta \in (0, \frac{1}{2}]$ and $\lambda \in (0, 1]$ be such that the following is satisfied:*

$$\sqrt{\ell' \lambda} \geq 50 \log(2\eta^{-1}) \sqrt{\max(\mathcal{N}(\lambda), 1)},$$

with ℓ' chosen to be the minimum of ℓ_2, ℓ_3, ℓ_4 from Lemma 2.4.1. Then, with probability at least $1-\eta$, the following holds:

$$\left\|(T_n + \lambda)^{-1}(T + \lambda)\right\| \leq 2.$$

Proof of Lemma 2.6.2 By means of the Neumann series decomposition we write:

$$(T_n + \lambda)^{-1}(T + \lambda) = (I - \Delta_\lambda)^{-1} = \sum_{j=0}^{\infty} \Delta_\lambda^j,$$

with $\Delta_\lambda := (T + \lambda)^{-1}(T - T_n)$. If $\|T_n(\lambda)\| < 1$, then the last series converges and the norm of $(T_n + \lambda)^{-1}(T + \lambda)$ is bounded by the sum of the series of norms. From Lemma 2.4.1, we have:

$$\|\Delta_\lambda\| \leq C_\eta \left(\sqrt{\frac{\mathcal{N}(\lambda)}{\lambda \ell'}} + \frac{2}{\lambda \ell'}\right),$$

where we put $C_\eta = 21 \log(2\eta^{-1})$ for $\eta \in (0, \frac{1}{2}]$. Using the lemma's assumption and the fact that $C_\eta > 28$ for $\eta \in (0, \frac{1}{2}]$, we obtain:

$$\sqrt{\lambda \ell'} \geq 2.3 C_\eta \sqrt{\max(\mathcal{N}(\lambda), 1)} \geq 2.3 C_\eta \geq 60.$$

This implies that

$$\frac{1}{\lambda \ell'} \leq \frac{1}{60\sqrt{\lambda \ell'}} \leq \frac{1}{120 C_\eta}.$$

Putting these pieces together we obtain:

$$\|\Delta_\lambda\| \leq C_\eta \left(\frac{1}{2.3 C_\eta} + \frac{1}{60 C_\eta}\right) < \frac{1}{2}.$$

This implies, that with probability at least $1-\eta$:

$$\left\|(T_n + \lambda)^{-1}(T + \lambda)\right\| \leq 2.$$

□

Proof [Sketch of the proof of Lemma 2.4.3] The proof is analogous in form and spirit to that of Propo-

sition 5.8 for the i.i.d. case given in Blanchard and Mücke (2018). The main difference is reflected in the use of the high probability upper bounds from Lemmata 2.4.1 and 2.6.2 instead of their i.i.d. counterparts, which in each case involve the knowledge of bounds on the effective sample size $\ell^{'}$. The appropriate choice of the latter is assured by the two conditions from the theorem statement. Namely, $\ell^{'} \geq \ell_0$ implies the claim of Lemma 2.6.2 (which is the $\tau-$mixing counterpart of the Lemma 5.4 from Blanchard and Mücke (2018)). On the other hand, the condition $\ell^{'} \leq \min\{\ell_2, \ell_3, \ell_4\}$ implies that all inequalities from Lemma 2.4.1 hold for $\ell^{'}$. We check additionally that the assumption $f_\nu \in \Omega(r, D)$ implies $\|f_\nu\| \leq D$ (since $\|T\| \leq 1$), which was a required condition for applying Lemma 2.4.1. The remaining reasoning is the same as in Proposition (5.8) from Blanchard and Mücke (2018).

□

Proof [Proof of Theorem 2.4.4] The proof of the first part of the Theorem is in essense a direct extension of the proof of Corollary 5.9 in Blanchard and Mücke (2018) to the case of $\tau-$mixing stationary sequence.

As the marginal distribution μ belongs to the class $\mathcal{P}^{<}(b, \beta)$ (by assumption), from Proposition 3 in De Vito et al. (2006), for any choice of parameter $\lambda \in (0, 1]$ we obtain:

$$\mathcal{N}(\lambda) \leq \tilde{C}_{b,\beta} \lambda^{-\frac{1}{b}}. \tag{2.34}$$

For the choice λ_n and $\ell_g^{'}$ given by (2.13) as function of n (the other parameters being fixed) it is easy to check by straightforward calculation that $\ell_g^{'} \geq \ell_0$ holds, where ℓ_0 is defined as in Lemma 2.4.3, provided n is larger than some n_0 (depending on all the fixed parameters).

Thus, as the given quantity $\ell_g^{'}$ fullfills all the requirements of Lemma 2.4.3, from this result we have with probability at least $1 - \eta$:

$$\left\| T^s \left(f_\nu - f_{\mathcal{D}_n}^{\lambda_n} \right) \right\|_{\mathcal{H}_k} \leq \tilde{C} \log(8\eta^{-1}) \lambda_n^s \left(D \left(\lambda_n^r + \frac{1}{\sqrt{\ell_g^{'}}} \right) + \frac{R}{\ell_g^{'} \lambda_n} + \sqrt{\frac{\Sigma^2 \lambda_n^{-\frac{b+1}{b}}}{\ell_g^{'}}} \right),$$

where $\tilde{C} := C_{r,s,b,\beta,\overline{\gamma},E,B,\chi,\gamma}$ depends potentially on all model and method parameters except for R, D and Σ.

By direct computation, we check that the choice of regularization parameter sequence λ_n implies that $\left(\ell_g^{'}\right)^{-1/2} = o_n(\lambda_n^r)$. Therefore, for n and $\ell_g^{'}$ large enough, we can disregard the term $\left(\ell_g^{'}\right)^{-1/2}$ in the above bound, by multiplying the bound in the front factor by 2. In the same vein, we can check that $(\ell_g^{'} \lambda_n)^{-1} = o_n\left(\left(\ell_g^{'}\right)^{-1/2} \lambda_n^{-\frac{b+1}{2b}}\right)$ and disregard the $\frac{R}{(\ell_g^{'} \lambda_n)}$ term provided n big enough. Finally, the proposed choice of parameter λ_n balances precisely the last two terms and leads to the conclusion.

□

Proof [Proof of Theorem 2.4.5] In this proof $C_\triangle$ will denote a factor depending on the model and method parameters (but not on n or η) whose exact value can change from line to line.

Observe that estimate (2.34) still holds, and additionally due to the assumption of lower bounded spectrum, a matching lower bound for the effective dimension holds (with a different factor). By relegating the effects of the constants R, K in the formulas from Table 2.1 in a generic factor, the choice of the bound for effective sample size $\ell_p^{'} = C_\triangle (\lambda_n \mathcal{N}(\lambda_n))^{\frac{2}{2\gamma+1}} n^{\frac{2\gamma}{2\gamma+1}}$ ensures that condition $\ell^{'} \leq \min\{\ell_2, \ell_3, \ell_4\}$ is fullfilled with $\ell^{'} = \ell_p^{'}$ (which can be checked by straightforward computation) and λ_n as defined by (2.15). Furthermore, for $n > n_0$, where n_0 is as specified in the statement of the theorem, we obtain :

$$\log \eta^{-1} \leq C_\triangle n^{\frac{br}{2br+b+1+b(r+1)\gamma^{-1}}},$$

which, by plugging in the value for λ_n and estimate for $\mathcal{N}(\lambda_n)$, implies that $\ell^{'} \geq \ell_0$; therefore we can apply Lemma 2.4.3. We get with probability at least $1-\eta$:

$$\begin{aligned}
&\left\| T^s\left(f_\nu - f_{\mathcal{D}_n}^{\lambda_n}\right)\right\|_{\mathcal{H}_k} \\
&\leq C_{\triangle,\eta}\lambda_n^s\left(\left(\lambda_n^r + \frac{\lambda_n^{-\left(\frac{b-1}{2b}\right)\frac{1}{2\gamma+1}}}{n^{\frac{\gamma}{2\gamma+1}}}\right) + \frac{1}{\lambda_n^{1+\frac{1}{2\gamma+1}\left(\frac{b-1}{b}\right)}n^{\frac{2\gamma}{2\gamma+1}}} + \lambda_n^{-\frac{1}{b}\frac{\gamma(b+1)+b}{2\gamma+1}}n^{-\frac{\gamma}{2\gamma+1}}\right)
\end{aligned}$$

where $C_{\triangle,\eta} = C_\triangle \log(8\eta^{-1})$. We observe that the choice of regularization parameter λ_n implies that $\lambda_n^{-\left(\frac{b-1}{2b}\right)\frac{1}{2\gamma+1}}/n^{\frac{\gamma}{2\gamma+1}} = o(\lambda_n^r)$. Therefore, similar to the case of exponentially $\tau-$mixing processes, assuming number of observations n large enough and multiplying the front factor with 2, we can disregard the term $\lambda_n^{-\left(\frac{b-1}{2b}\right)\frac{1}{2\gamma+1}}/n^{\frac{\gamma}{2\gamma+1}}$ in the above bound. Similarly, one can check that:

$$\frac{1}{\lambda_n^{1+\frac{1}{2\gamma+1}\left(\frac{b-1}{b}\right)}n^{\frac{2\gamma}{2\gamma+1}}} = o\left(\lambda_n^{-\frac{1}{b}\frac{\gamma(b+1)+b}{2\gamma+1}}n^{-\frac{\gamma}{2\gamma+1}}\right),$$

so this term can be similarly asymptotically disregarded (again, multiplying the second term by 2). Therefore, we can concentrate the analysis on the remaining main terms which are λ_n^r and $\lambda_n^{-\frac{1}{b}\frac{\gamma(b+1)+b}{2\gamma+1}}n^{-\frac{\gamma}{2\gamma+1}}$. The choice of λ_n balances exactly these terms and the computations lead to the conclusion. □

Chapter 3

Online nonparametric regression with kernels

Contents

In this part of the thesis we investigate the variation of the online kernelized ridge regression algorithm in the setting of $d-$dimensional adversarial nonparametric regression. We derive the regret upper bounds on the classes of Sobolev spaces $W_p^\beta(\mathcal{X})$, $p \geq 2, \beta > \frac{d}{p}$. The upper bounds are supported by the minimax regret analysis, which reveals that in the cases $\beta > \frac{d}{2}$ or $p = \infty$ these rates are (essentially) optimal. Finally, we compare the performance of the kernelized ridge regression forecaster to the known nonparametric forecasters in terms of the regret rates and their computational complexity as well as to the excess risk rates in the setting of (i.i.d.) nonparametric regression. This chapter is based on the joint work with Pierre Gaillard, Sebastien Gerschinovitz, Alessandro Rudi, which can be found in Zadorozhnyi et al. (2021).

3.1 Introduction

In this chapter, we consider the online least-squares regression framework (Cesa-Bianchi and Lugosi, 2006) as a game between the environment and the learner where the task is to sequentially predict the environment's output y_t given the current input x_t and the observed history $\{(x_i, y_i)\}_{i=1}^{t-1}$. Specifically, let $\mathcal{X} \subset \mathbb{R}^d$ be an input space, $\mathcal{Y} \subset \mathbb{R}$ a label space, and $\widehat{\mathcal{Y}} \subset \mathbb{R}$ a target space. Before the game starts, the environment secretly produces a sequence of input–output pairs $(x_1, y_1), (x_2, y_2), \ldots$ in $\mathcal{X} \times \mathcal{Y}$ over some (possibly infinite) time horizon.

At each round $t \geq 1$, the environment first reveals an input $x_t \in \mathcal{X}$; the learner forms the prediction $\widehat{y}_t \in \widehat{\mathcal{Y}}$ of the true label $y_t \in \mathcal{Y}$ based on past information $(x_1, y_1), \ldots, (x_{t-1}, y_{t-1}) \in \mathcal{X} \times \mathcal{Y}$ and on the current input x_t. The true label y_t is then revealed, the learner suffers the squared loss $(y_t - \widehat{y}_t)^2$ and round $t + 1$ starts. The problem is to design an algorithm which minimizes the learner's cumulative regret

$$R_n(\mathcal{F}) := \sup_{f \in \mathcal{F}} R_n(f)\,, \qquad \text{where} \quad R_n(f) := \sum_{t=1}^{n} (y_t - \widehat{y}_t)^2 - \sum_{t=1}^{n} (y_t - f(x_t))^2\,, \tag{3.1}$$

over $n \geq 1$ rounds with respect to the best prediction rule from some reference functional class $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$.

In the setting of adversarial online learning the nature of data can be completely arbitrary, unlike in the standard statistical learning framework where the data stream is assumed to be generated from some underlying stochastic process, usually with an independent noise component. The problem of online learning with arbitrary (adversarial) data goes back to the work of Foster (1991). Much theoretical research has been done since then for parametric models (e.g. Azoury and Warmuth, 2001; Cesa-Bianchi, 1999b; Vovk, 1998). However, the amount of data and the complexity of current machine learning problems have led the community to explore the more general problem of online-learning with methods based on nonparametric decision rules and with the reference classes being bounded functional sets (see ex. Vovk (2006a), Rakhlin and Sridharan (2014)). Much effort has been devoted to the regret analysis with respect to functional classes that include Sobolev spaces (Rakhlin and Sridharan, 2014; Rakhlin et al., 2014; Vovk, 2006a, 2007). Surprisingly, only a few explicit algorithms have been designed to address the regression problem (Gaillard and Gerchinovitz, 2015; Vovk, 2006a,b, 2007). Although they have optimal (or close to optimal) regret rates, these algorithms have the disadvantage of either being computationally intractable or of providing suboptimal regret upper bounds (see Table 3.1 for computational complexities of some known algorithms). For more details on previous work, we refer the reader to Section 3.5.

In this work we consider the framework of online adversarial regression where the benchmark class $\mathcal{F}$, against which the algorithm competes, is a ball in a Sobolev space (see e.g., Adams and Fournier, 2003), denoted by $W_p^\beta(\mathcal{X})$ where $\beta > 0$ and $p \geq 2$. In other words, $\mathcal{F}$ is the space of functions with p-integrable weak derivatives up to order β (see 3.2.2 for more details).

The problem is of interest since, to date, the optimal regret is achieved only by a computationally efficient algorithm in the smooth regime when $\beta > d/2$ and $p = 2$.

Overview of the main results and outline of the chapter We provide an in-depth analysis of the regret achieved by a version of the online kernel ridge regression algorithm, Kernel Aggregating Algorithm Regression (KAAR) over the classes of bounded balls in $W_p^\beta(\mathcal{X})$. In particular the key contribution is the analysis of the robustness of the KAAR which returns an element in RKHS while competing against a function which does not belong to a RKHS. We notice that (on the contrary to many known nonparametric schemes, see for example Rakhlin and Sridharan (2014), Vovk (2007)) this algorithm is computationally tractable. Comparison of the performance of KAAR to the known procedures (both in regret rates and computational efficiency) is summarized in Table 3.1. Furthermore, we also prove lower

	KAAR (3.10)		Rakhlin and Sridharan (2014)		Chaining EWA Gaillard and Gerchinovitz (2015)			EWA by Vovk (2006a)	
	Regret[1]	Cost	Regret	Cost	Regret	Cost	Cost ($d=1, p=\infty$)[2]	Regret	Cost
$\beta > \frac{d}{2}$	$n^{1-\frac{2\beta}{2\beta+d}+\varepsilon}$	$n^3 + dn^2$	$n^{1-\frac{2\beta}{2\beta+d}}$	Non constructive	$n^{1-\frac{2\beta}{2\beta+d}}$	$\exp(n)$	poly(n)	$n^{1-\frac{\beta}{\beta+d}}$	$\exp(n)+nd$
$\frac{d}{p} < \beta \leq \frac{d}{2}$	$n^{1-\frac{\beta}{d}\frac{p-d/\beta}{p-2}+\varepsilon}$	$n^3 + dn^2$	$n^{1-\frac{\beta}{d}}$	Non constructive	$n^{1-\frac{\beta}{d}}$	$\exp(n)$	$n^{\lceil\beta\rceil\left(\frac{5\beta+2}{2\beta+1}\right)}$	$n^{1-\frac{\beta}{\beta+d}}$	$\exp(n)+nd$
$p=\infty, \beta \leq \frac{d}{2}$	$n^{1-\frac{\beta}{d}+\varepsilon}$	$n^3 + dn^2$	$n^{1-\frac{\beta}{d}}$	Non constructive	$n^{1-\frac{\beta}{d}}$	$\exp(n)$	$n^{\lceil\beta\rceil\left(\frac{5\beta+2}{2\beta+1}\right)}$	$n^{1-\frac{\beta}{\beta+d}}$	$\exp(n)+nd$

Table 3.1: Regret rates and time complexity of KAAR (3.10) (new upper-bounds from are highlighted in blue) and the existing algorithms for online nonparametric regression.

bounds for minimax regret (which is defined as the infimum over all admissible strategies of a supremum of all data-sequences) reaches optimal or close to optimal (up to a polynomial factor in the number of rounds) regret rates on bounded balls of Sobolev spaces $W_p^\beta(\mathcal{X})$ with $p \geq 2$ and $\beta > \frac{d}{p}$ (which, up to a multiplicative constant which depends on the diameter of the set, implies the result for all bounded subsets of continuous functions in $W_p^\beta(\mathcal{X})$).

More precisely, the result is threefold. On the one hand, our analysis recovers the classical result for Sobolev spaces, i.e. when $\beta > d/2$ and $p \geq 2$. In particular, we show in Theorem 3.3.2 that, choosing appropriate regularization parameter, on the classes of continuous functions which belong to Sobolev RKHS of smoothness β, KAAR achieves the optimal regret upper bound[3]

$$R_n(\mathcal{F}) \lesssim n^{1-\frac{2\beta}{2\beta+d}} \log n.$$

On the other hand, we consider the more challenging scenario when $d/2 > \beta \geq d/p$, which corresponds less smooth benchmark functional classes that cannot be embedded into a RKHS. We will refer to this case as the *hard-learning scenario*. In Theorem 3.3.4 we prove that in such a scenario, when $\mathcal{F} = B_{W_p^\beta(\mathcal{X})}(0, R)$, the regret of KAAR with well-chosen parameters is upper-bounded as

$$R_n(\mathcal{F}) \lesssim n^{1-\frac{\beta}{d}\frac{p-\frac{d}{\beta}}{p-2}} \log n.$$

In particular, when $p = \infty$, the regret upper bound is of order $O(n^{1-\frac{\beta}{d}+\varepsilon} \log n)$. The latter bound is then proven to be optimal (up to a constant ε that can be made arbitrary small) by showing the lower bound for minimax regret in Section 3.4 for the lower bounds. Optimal regret upper bounds on the classes of bounded Hölder balls were previously derived with polynomial-time algorithms for $d = 1$ Gaillard and Gerchinovitz (2015). The case $d \geq 1$ and $\beta = 1$ was also analyzed for Lipschitz and semi-Lipschitz losses in Cesa-Bianchi et al. (2017). Notice that throughout the chapter we do not consider the case of Sobolev spaces with $\beta \leq d/p$. In the latter case, the existence of continuous representatives for equivalence classes in $W_p^\beta(\mathcal{X})$ is not guaranteed, and the regret of any forecaster will be linear.

In Figure 3.1, we plot the regions of the $(1/p, \frac{\beta}{d})$-plane corresponding to the different regret cases where we obtain either the optimal rate or a suboptimal rate, that nevertheless improved with respect to classical aggregation algorithms in the nonparametric framework Vovk (2006a). Note that the smaller $\frac{\beta}{d}$ and p are, the harder the problem is. Additional graphs comparing the regret of KAAR with the EWA forecaster are available in Section 3.6.7.

To complete the analysis of online nonparametric regression over Sobolev spaces, we make use of the general results of Rakhlin and Sridharan (2014), derive upper and lower bound on the fat-shattering

[1] In terms of its upper bound.

[2] Gaillard and Gerchinovitz (2015) only provide an efficient version of their algorithm for Sobolev spaces with $p = \infty$, $d = 1$ and $\beta \geq 1/2$. Their efficient algorithm can however be extended for any $\beta \in (0, 1/2)$ with a polynomial time complexity.

[3] The notation $\lesssim$ denotes an approximate inequality which includes multiplicative constants which depend on $\mathcal{F}$ and $\mathcal{X}$.

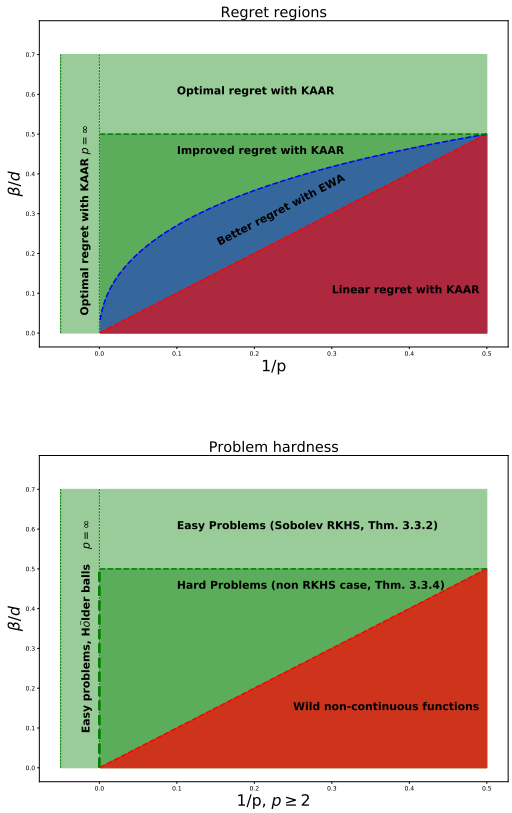

Figure 3.1: (Left) Different regions in the $(1/p, \frac{\beta}{d})$-plane for which our new regret bound for KAAR: [light green] is optimal (i.e., $\beta > d/2$ or $p = \infty$); [dark green] improves the bound of EWA by Vovk (2006b); [blue] is worse than the bound of EWA; [red] is linear in n (i.e, $\beta \leq d/p$). (Right) Hardness of the problem in the $(1/p, \frac{\beta}{d})$ plane

dimension (see Rakhlin and Sridharan (2014) and 3.6.6 for an exact definition) and establish corresponding lower bounds. We prove that any admissible algorithm (– the exact definition of which will be presented in section 3.4) suffers at least the minimax regret of order $n^{1-2\beta/(2\beta+d)}$ in the smooth case $\beta > d/2$, and $n^{1-\beta/d}$ when $\beta \leq d/2$. The latter implies that KAAR (with the proper choice of parameters) achieves optimal regret rates when $\beta > d/2$ or $p = \infty$. The regret analysis of KAAR on the classes of compact subsets of Sobolev spaces $W_p^\beta(\mathcal{X})$ and $p \geq 2$ as well as lower bounds for minimax regret for the classes of bounded balls in Sobolev spaces $W_p^\beta(\mathcal{X})$ are summarized in Table 3.2.

	Upper bound of KAAR	Lower bound for minimax regret
$\beta > \frac{d}{2}$	$n^{1-\frac{2\beta}{2\beta+d}+\varepsilon}\log(n)$	$n^{1-\frac{2\beta}{2\beta+d}}$
$\frac{d}{p} < \beta \leq \frac{d}{2}$	$n^{1-\frac{\beta}{d}\frac{p-d/\beta}{p-2}+\varepsilon}\log(n)$	$n^{1-\frac{\beta}{d}}$
$p = \infty, \beta \leq \frac{d}{2}$	$n^{1-\frac{\beta}{d}+\varepsilon}\log(n)$	$n^{1-\frac{\beta}{d}}$

Table 3.2: Regret upper bounds of KAAR and the corresponding lower bound on the classes of bounded subsets of $W_p^\beta(\mathcal{X})$, $\beta \in \mathbb{R}$, $p \geq 2$. Here $\varepsilon > 0$ is an arbitrary small number.

The outline of the rest of the chapter is constituted as follows. In Section 3.2, we fix the notation

and recall the definition of Sobolev spaces, reproducing kernel Hilbert spaces (RKHS) and their effective dimension. Furthermore, we describe KAAR therein. In Section 3.3, we provide our regret upper bounds for KAAR and in Section 3.4 we present the corresponding lower bounds. Finally, in Section 3.5, we make more detailed comparisons with existing work both in the adversarial online regression setting studied in this chapter and in the more standard statistical framework with i.i.d. observations. We discuss the optimality of the rates and comment on the aspect of computational complexity by showing that KAAR is superior to the known nonparametric schemes in terms of runtime and storage complexities. All the proofs as well as technical details on Sobolev spaces and kernels are given in the Appendices.

3.2 Notation and background

3.2.1 Kernels and effective dimension

We recall below some notations on reproducing kernel Hilbert Spaces (RKHS) which is also used in the Chapter 2 and for the setup of the usage of kernel methods we refer to Chapter1.

We consider a real-valued kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ and the corresponding reproducing kernel Hilbert space, which we denote $\mathcal{H}_k$ and the (canonical) feature map $k_x := k(x, \cdot)$. The prediction rule f_t at round t then forecasts $\widehat{f}_t(x_t) = \langle f_t, k_{x_t} \rangle_{\mathcal{H}_k}$. We recall the following notations which were used in Chapter 2.

Namely, for $t \geq 1$ and a data sample $\mathcal{D} = \{x_s, y_s\}_{s=1}^t$, RKHS $\mathcal{H}_k$ which is generated by a kernel k, $f \in \mathcal{H}_k$ and $y \in \mathbb{R}^t$ we defined $S_t : \mathcal{H}_k \mapsto \mathbb{R}^t$, $S_t f = (f(x_1), \ldots, f(x_t)) \in \mathbb{R}^t$, $S_t^* : \mathbb{R}^t \mapsto \mathcal{H}_k$, $S_t^* y = \sum_{s=1}^t y_s k_{x_s}$, $T_t : \mathcal{H}_k \mapsto \mathcal{H}_k$, $T_t f = \sum_{s=1}^t k_{x_s} \langle f, k_{x_s} \rangle$. Lastly, we recall the notion of the effective dimension.

Definition 3.2.1. Effective dimension For a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, datasample $\mathcal{D}_n = \{x_i\}_{1 \leq i \leq n} \in \mathcal{X}^n$ and $\tau > 0$ the effective dimension of RKHS $\mathcal{H}_k$ associated with the sample $\mathcal{D}_n$ on the scale τ is given as

$$d_{eff}^n(\tau) := \mathrm{Tr}\left((K_n + \tau I)^{-1} K_n\right) = \sum_{j=1}^n \frac{\lambda_j(K_n)}{\lambda_j(K_n) + \tau}, \tag{3.2}$$

where $I : \mathbb{R}^n \mapsto \mathbb{R}^n$ is the identity matrix.

In statistical learning, it has been shown (Zhang (2005), Rudi et al. (2015), and Blanchard and Mücke (2018)) that the effective dimension characterizes the generalization error of kernel-based algorithms. This is a decreasing function of the scale parameter τ and $d_{eff}^n(\tau) \to 0$ when $\tau \to \infty$. On the other side, as $\tau \to 0$, it converges to the rank of K_n, which can be interpreted as the "physical" dimension of the points $(k_{x_i})_{1 \leq i \leq n}$.

3.2.2 Sobolev Spaces

Let $\beta \in \mathbb{N}_*$, $2 \leq p < \infty$ and $\mathcal{X} := [-1, 1]^d$, where we use standard notation for $\mathbb{N}_* := \{1, 2 \ldots, \}$. We denote by $L_p(\mathcal{X})$ the space of equivalence classes of p-integrable functions with respect to the Lebesgue measure λ on the Borel $\sigma-$algebra $B(\mathcal{X})$ and by $[f]_\lambda$ the $\lambda-$equivalence class to some function $f : \mathcal{X} \mapsto \mathbb{R}$. We denote by $C^m(\mathcal{X})$ the space of all m-times differentiable functions f with multidimensional derivative $D^\gamma f$ ($|\gamma|_1 \leq m$) that are continuous on $\mathcal{X}$ and let $C(\mathcal{X})$ be the standard space of continuous functions equipped with the norm $\|f\|_{C(\mathcal{X})} = \max_{x \in \mathcal{X}} |f(x)|$ (we write it simply $\|f\|$ when no confusion can arise). For the normed space $(\mathcal{G}, \|\cdot\|)$ we use $B_{\mathcal{G}}(x, R)$ and $\overline{B}_{\mathcal{G}}(x, R)$ to denote respectively the open and the closed ball of radius R centered at the point x.

Definition of Sobolev spaces. We denote $|\gamma|_1 := \sum_{i=1}^n |\gamma_i|$ for $\gamma \in \mathbb{N}_*^d$ and we write $D^\gamma f$ for the multidimensional weak derivative (see section 5.2.1, page 242 in **?**) of the function $f : \mathcal{X} \mapsto \mathbb{R}$ of order

$\gamma \in \mathbb{N}_*^d$. We recall that the Sobolev space (see chapter Adams and Fournier (2003)) $W_p^\beta(\mathcal{X})$ is the space of all equivalence classes of functions $[f]_\lambda \in L_p(\mathcal{X})$ such that

$$\|f\|_{W_p^\beta(\mathcal{X})} := \begin{cases} \left(\sum_{|\gamma|_1 \le \beta} \|D^\gamma f\|_{L_p(\mathcal{X})}^p\right)^{\frac{1}{p}} & \text{if} \quad p < \infty \\ \sup_{|\gamma|_1 \le \beta} \|D^\gamma f\|_{L_\infty(\mathcal{X})} & \text{if} \quad p = \infty \end{cases}$$

is finite. The notion of Sobolev spaces is then extended to the case of any real $\beta > 0$ by means of the Gagliardo (semi)norms. In the case $p = 2$ it can be shown to be equivalent to the known approach of the definition of fractional Sobolev spaces via Fourier transform. Let $\mathcal{X} \subseteq \mathbb{R}^d$, $p \in [1, \infty)$ and denote $L_p(\mathcal{X})$ for the equivalence class of $p-$integrable functions with respect to the Lebesque measure λ on $\mathcal{X}$. We firstly recall (see Adams and Fournier (2003), chapter 3) the definition of Sobolev spaces with integer exponent. Denote $L_p(\mathcal{X})$ for the equivalence class of $p-$integrable functions with respect to the Lebesque measure λ on $\mathcal{X}$. Classes $W_p^r(\mathcal{X})$ and $W_\infty^r(\mathcal{X})$ are the vector spaces of equivalence classes of functions defined as:

$$W_p^r(\mathcal{X}) := \left\{ f : \mathcal{X} \to \mathbb{R} \quad \text{s.t.} \quad \|f\|_{W_p^r(\mathcal{X})} := \Big(\sum_{|\gamma|_1 \le r} \|D^\gamma f\|_{L_p(\mathcal{X})}^p \Big)^{\frac{1}{p}} < \infty \right\},$$

and

$$W_\infty^r(\mathcal{X}) := \left\{ f : \mathcal{X} \to \mathbb{R} \quad \text{s.t.} \quad \|f\|_{W_\infty^r(\mathcal{X})} := \sup_{|\gamma|_1 \le r} \|D^\gamma f\|_{L_\infty(\mathcal{X})} < \infty \right\}.$$

We also define the Sobolev semi-norm $|f|_{W_p^j(\mathcal{X})} := \sum_{\gamma:|\gamma|=j} \|D^\gamma f\|_{L_p(\mathcal{X})}$. Now, for $\beta \in \mathbb{R}_+$ write $\beta = r + \sigma$ with $r \in \mathbb{N}_0$ and $\sigma \in (0, 1)$, i.e. $r = \lfloor \beta \rfloor$, $\sigma = \beta - \lfloor \beta \rfloor$. Let $u : \mathcal{X} \mapsto \mathbb{R}$ be some fixed measurable function. We define the map $\varphi_u : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R} \cup \{\infty\}$ such that for $1 \le p < \infty$ and all $(x, y) \in \mathcal{X} \times \mathcal{X}$:

$$\varphi_u(x, y) = \frac{|u(x) - u(y)|}{\|x - y\|_2^{\frac{d}{p} + \sigma}},$$

and denote

$$\tilde{W}_p^\sigma(\mathcal{X}) := \{u \in L_p(\mathcal{X}) : \|\varphi_u\|_{L_p(\mathcal{X} \times \mathcal{X})} < \infty\}.$$

The space $\tilde{W}_p^\sigma(\mathcal{X})$ equipped with the norm $\|u\|_{\tilde{W}_p^\sigma(\mathcal{X})} := \left(\|u\|_{L_p(\mathcal{X})} + \|\varphi_u\|_{L_p(\mathcal{X} \times \mathcal{X})}\right)^{\frac{1}{p}}$ can be shown to be a Banach space. With this notation, Sobolev space $W_p^\beta(\mathcal{X})$, $\beta = r + \sigma$ can be defined as

$$W_p^\beta(\mathcal{X}) := \left\{ u \in W_p^r(\mathcal{X}) : D^\gamma u \in \tilde{W}_p^\sigma(\mathcal{X}) \text{ for any } \gamma \text{ such that } |\gamma|_1 = r \right\}. \tag{3.3}$$

Equipped with the norm

$$\|u\|_{W_p^\beta(\mathcal{X})} := \left(\|u\|_{W_p^r(\mathcal{X})}^p + \sum_{\gamma:|\gamma|=r} \|D^\gamma u\|_{\tilde{W}_p^\sigma(\mathcal{X})}^p \right)^{\frac{1}{p}}, \tag{3.4}$$

it becomes Banach space. In the case $\beta = m \in \mathbb{N}_*$, it matches the definition of the Sobolev space $W_p^m(\mathcal{X})$ (up to a re-scaling of the norm). If $m = 0$ (i.e. $r = \sigma \in [0, 1)$), we find that $W_p^m(\mathcal{X}) = L_p(\mathcal{X})$ so that the norm in $W_p^\sigma(\mathcal{X})$ is given by

$$\|u\|_{W_p^r(\mathcal{X})} = \|u\|_{\tilde{W}_p^\sigma(\mathcal{X})} := \left(\|u\|_{L_p(\mathcal{X})} + \|\varphi_u\|_{L_p(\mathcal{X} \times \mathcal{X})}\right)^{\frac{1}{p}}. \tag{3.5}$$

In accordance with the above definition of the class $W_p^\beta(\mathcal{X})$, for any $\beta = r + \sigma$, $\sigma \in [0, 1)$ we set

$$\tilde{W}_\infty^\sigma(\mathcal{X}) := \{u \in L_\infty(\mathcal{X}) : \sup_{x,y\in\mathcal{X}, x\neq y} \frac{|u(x) - u(y)|}{\|x - y\|_2^\sigma} \leq \infty\}. \tag{3.6}$$

Now, for $\beta = r + \sigma \in \mathbb{R}$ the Sobolev space $W_\infty^\beta(\mathcal{X})$ can be defined as a functional space

$$W_\infty^\beta(\mathcal{X}) := \left\{u \in W_\infty^m(\mathcal{X}) : D^\gamma u \in \tilde{W}_\infty^\sigma(\mathcal{X}) \text{ for any } \gamma \text{ such that } |\gamma|_1 = r \right\}. \tag{3.7}$$

equipped with a norm

$$\|u\|_{W_\infty^\beta(\mathcal{X})} := \max\{\|u\|_{W_\infty^r(\mathcal{X})}, \max_{\gamma:|\gamma|_1=r} \|D^r u\|_{\tilde{W}_\infty^\sigma(\mathcal{X})}\} \tag{3.8}$$

Sobolev Reproducing Kernel Hilbert Spaces. We recall here known results on embedding characteristics of fractional Sobolev Hilbert spaces, which are essential in our analysis. Let $s \in \mathbb{R}_+$, and consider the Sobolev space $W_2^s(\mathcal{X})$ with $\mathcal{X} \subset \mathbb{R}^d$. It is a separable Hilbert space (see Chapter 7 in Schaback (2007)) with the inner product $\langle f, g\rangle = \sum_{\|\gamma\|_1 \leq s} \langle D^\gamma f, D^\gamma g\rangle_{L_2(\mathcal{X})}$. By the Sobolev Embedding Theorem (see Theorem 7.34 in Adams and Fournier (2003) for the case $s \in \mathbb{R}_+$, $s > d/2$) we have that $W_2^s(\mathcal{X}) \hookrightarrow C(\mathcal{X})$. The latter embedding is to be understood in the sense that there exists $C_1 > 0$, such that each $\lambda-$equivalence class has a unique element $f \in C(\mathcal{X})$ such that $\|f\|_{C(\mathcal{X})} \leq C_1 \|f\|_{W_2^s(\mathcal{X})}$. We refer to the set of continuous representatives of all equivalence classes in $W_2^s(\mathcal{X})$ as to *Sobolev RKHS* and denote it as $W^s(\mathcal{X})$. It can be shown (see paragraph 7.5 and Theorem 7.13 in Schaback (2007)) that $W^s(\mathcal{X})$ is a RKHS. Furthermore (see part (c) Theorem 7.34 in Adams and Fournier (2003)), when $p \geq 2$, $W_p^s(\mathcal{X})$ is embedded into the space of continuous functions $C(\mathcal{X})$ if $s > d/p$ while if $s < \frac{d}{2}$ is not (and not embeddable into) a RKHS.

Furthermore, (see chapter 7 in Schaback (2007)), Sobolev RKHS $W^s(\mathcal{X})$ is generated by the translation invariant kernel, which is a restriction to $\mathcal{X}$ of the kernel k_s of $W^s\left(\mathbb{R}^d\right)$ (see also Corollary 10.48 on page 170 in Wendlandt (2005)). It is a continuous, bounded and measurable kernel (see general Lemma 4.28 and 4.25 in Steinwart and Christmann (2008) –) which is defined for all $x, x^{'} \in \mathcal{X}$ by

$$k\big(x, x'\big) := \frac{2^{1-s}}{\Gamma(s)} \big\|x - x'\big\|_2^{s-\frac{d}{2}} K_{\frac{d}{2}-s}\big(\big\|x - x'\big\|_2\big)\,, \tag{3.9}$$

where $K_{d/2-s}(\cdot)$ is a modified Bessel function of the second kind (see Chapter 5.1 in Wendlandt (2005) for more details on Bessel function). Alternatively, the kernel function $k(\cdot)$ of Sobolev RKHS $W^s(\mathcal{X})$ can be described by its Fourier transform, which equals $\mathcal{F}(k)(\omega) = (1 + \|\omega\|_2^2)^{-s}$. We refer the reader to Chapters 10–11 in Wendlandt (2005) as well as to Novak et al. (2017) for more details on the kernel functions of Sobolev RKHS.

3.2.3 Main Algorithm — KAAR

In this part we analyse the regret achieved by KAAR (Gammerman et al. (2004)), over the (Sobolev) RKHS $W^s(\mathcal{X})$. The regret is measured with respect to the benchmark classes of bounded Sobolev balls $W_p^\beta(\mathcal{X})$ which may have different regularity (i.e. we consider the case when $\beta \neq s$).

KAAR (see Algorithm 1) was introduced in the case of adversarial sequential linear regression by Vovk (2001) and Azoury and Warmuth (2001); further it was analyzed in Cesa-Bianchi and Lugosi (2006), Rakhlin and Sridharan (2014), Gaillard et al. (2019) and applied to concrete forecasting problems including electricity (Devaine et al., 2013), air quality (Mallet et al., 2009) and exchange rate (Amat et al., 2018) forecasting. It was extended to the case of general reproducing Hilbert spaces in Gammerman et al. (2004), whereas Jézéquel et al. (2019) provide a variation of the algorithm with the same regret and

Parameters: $d \geq 1$, $s > d/2$, and $\tau > 0$
Initialization: define $k(\cdot, \cdot)$ as in (3.9);
while $t \geq 1$ **do**
  observe $x_t \in \mathcal{X}$;
  $\tilde{y}_t := (y_1, \ldots, y_{t-1}, 0)^\top$;
  $\tilde{k}(x_t) := \big(k(x_1, x_t), \ldots, k(x_{t-1}, x_t), k(x_t, x_t)\big)$;
  $K_t := \big(k(x_i, x_j)\big)_{1 \leq i,j \leq t}$;
  forecast $\widehat{y}_t := \tilde{y}_t^\top (K_t + \tau I_t)^{-1} \tilde{k}(x_t)$;
  observe y_t;
end

Algorithm 1: KAAR (Gammerman et al., 2004) on Sobolev RKHS

reduced computational complexity. In the case of Sobolev spaces, KAAR (Alg. 1) reads as follows. Let $\tau > 0$, $s > d/2$ at round $t \geq 1$ KAAR predicts $\widehat{y}_t := \widehat{f}_{\tau,t}(x_t)$, where

$$\widehat{f}_{\tau,t} := \operatorname*{Arg\,Min}_{f \in W^s(\mathcal{X})} \left\{ \left(\sum_{j=1}^{t-1} \big(y_j - f(x_j)\big)^2 \right) + \tau \|f\|^2_{W_2^s(\mathcal{X})} + f^2(x_t) \right\}. \tag{3.10}$$

The prediction $\widehat{y}_t = \widehat{f}_{\tau,t}(x_t)$ can be computed in the closed form by Algorithm 1 in $\mathcal{O}(n^3 + n^2 d)$ operations (see Section 3.5.3 for details on the computational complexity). This improves computational complexity over other known nonparametric online regression algorithms, which achieve optimal regret with respect to Sobolev spaces in dimension d.

Remark 3.2.2. We remark that the right–hand side of (3.10) depends on the input x_t, so while $\widehat{f}_t^\tau \in W^s(\mathcal{X})$, the prediction function $x_t \mapsto \widehat{f}_{\tau,t}(x_t)$ is a measurable function which in general not necessarily belongs to the space $W_2^s(\mathcal{X})$, the prediction map does not necessary belong to the benchmark class against which the algorithm is competing. This corresponds to the so-called case of improper learning (see more details in Rakhlin et al. (2015),**?**). Furthermore, a sequential version of kernel ridge regression was considered by Zhdanov and Kalnishkan (2010). It removes the term $f^2(x_t)$ in the r.h.s. of (3.10) and clips the prediction, by forecasting $\widehat{y}_t^M := \widehat{y}_t := \min(\max(-M, \tilde{f}_{\tau,t}(x_t)), M)$, where $\tilde{f}_t$ is the solution to the Problem 3.10 without $f^2(x_t)$ term. Since for every $y_t \in [-M, M]$ we have $\big(y_t - \widehat{y}_t^M\big)^2 \leq (y_t - \widehat{y}_t)^2$ so for the clipped version of the KAAR forecaster $\widehat{y}_t^M$, the upper bound regret analysis for KAAR directly applies to its clipped version (given the same, i.e. unclipped benchmark class).

We emphasize that throughout the chapter β and p refer to the parameters of the benchmark Sobolev space and $s > d/2$ refers to the smoothness parameter of RKHS $W^s(\mathcal{X})$ used in KAAR.

3.3 Main results: Upper-bound on the regret of KAAR on the classes of Sobolev balls.

In this section, we present regret upper bounds of KAAR on the reference classes of bounded balls in $W_p^\beta(\mathcal{X})$, $\beta > \frac{d}{p}$. By the Sobolev embedding Theorem (see Adams and Fournier (2003) Theorem 7.34 or Equation 10 on page 60 in Edmunds and Triebel (1996)), condition $\beta > \frac{d}{p}$ implies that every equivalence class in $W_p^\beta(\mathcal{X})$ has a continuous representative. In our analysis under $R_n\Big(B_{W_p^\beta(\mathcal{X})}(0, R)\Big)$ we always understand regret with respect to the correspondent ball of continuous representatives bounded in the norm of the space $W_p^\beta(\mathcal{X})$ (Adams and Fournier, 2003).

We consider the framework of online adversarial regression with the label space $\mathcal{Y} := [-M, M]$ real-valued predictions, the input space $\mathcal{X} = [-1, 1]^d$ and the reference class $\mathcal{F} := B_{W_p^\beta(\mathcal{X})}(0, R)$ being an

open ball in Sobolev space $W_p^\beta(\mathcal{X})$ of radius $R > 0$ with $d \in \mathbb{N}_*, \beta \in \mathbb{R}_+$ and $p \geq 2$, where we use standard notation for $\mathbb{N}_* = \{1, 2 \ldots, \}$. We remark that the assumption on the input space is given for simplicity and can be weakened to any bounded domain in $\mathbb{R}^d$ with Lipschitz boundary (see Chapter 4 in Adams and Fournier (2003) on more details on Lipschitz boundaries).

3.3.1 Key preliminary result and the upper-bound on the effective dimension.

We start by recalling a general upper bound on the regret of KAAR on the bounded balls of the general separable RKHS in terms of the effective dimension. It is a direct extension of the upper bound of KAAR in Vovk (2001) and Azoury and Warmuth (2001) from finite dimensional linear regression to kernel regression and can be retrieved from Theorem 2 in Gammerman et al. (2004) (see also Proposition 1 and 2 in Jézéquel et al. (2019) for the next statement) for the case of Sobolev RKHS, as the underlying kernel function is continuous. The regret of KAAR on any $f \in W_2^s(\mathcal{X})$ is upper-bounded as

$$R_n(f) \leq \tau \|f\|^2_{W_2^s(\mathcal{X})} + M^2 \left(1 + \log\left(1 + \frac{n\kappa^2}{\tau}\right)\right) d_{eff}^n(\tau), \tag{3.11}$$

where $\kappa > 0$ is such that $\sup_x k(x,x) \leq \kappa^2$ and $d_{eff}^n(\tau)$ is the effective dimension as given in Definition 3.2.1. The regret bound (3.11) will be used as a starting point to prove different upper-bounds in the next subsection.

Theorem 3.3.1 provides an upper bound on the effective dimension for the Sobolev RKHS $W^s(\mathcal{X})$.

Theorem 3.3.1 (Upper bound for the effective dimension of Sobolev RKHS)**.** *Let $\varepsilon \in (0, 1/4)$, $d \geq 1$, $n > 1$ and $s > d/2$. Consider the Sobolev RKHS $W^s(\mathcal{X})$ with $\mathcal{X} := [-1,1]^d$. For any sequence of inputs $\mathcal{D} := \{x_1, \ldots, x_n\}$ and $\tau > 0$, the effective dimension $d_{eff}(\tau)$ is upper-bounded as*

$$d_{eff}^n(\tau) \leq C\left(\left(\frac{n}{\tau}\right)^{\frac{d}{2s}+\varepsilon_1} + 1\right),$$

where $\varepsilon_1 = d\varepsilon/s^2$, and C[4] is a constant which depends on $d, s, R, K, M, \mathcal{X}, \varepsilon$ but is independent of n. Furthermore, if $s \in \mathbb{N}_*$, then $\varepsilon = 0$.

The proof of this statement, is presented in 3.6.3. It is based on some known properties of low rank projections in Sobolev spaces which are recalled in 3.6.1.

3.3.2 Regret upper bound for the Sobolev RKHS ($\beta > d/2$)

Notice that when $p \geq 2$ and $\beta \geq d/2$ we have $W_p^\beta(\mathcal{X}) \subseteq W_2^\beta(\mathcal{X})$ and (by the Sobolev embedding theorem) $W_p^\beta(\mathcal{X}) \mapsto C(\mathcal{X})$. Because the space of continuous representatives of every equivalence class $W_p^\beta(\mathcal{X})$ is a closed subspace of $W^s(\mathcal{X})$, it is a RKHS. Using KAAR with $s = \beta$ and putting the upper bound for the effective dimension of $W^\beta(\mathcal{X})$ into the regret upper bound (3.11) with the proper choice of the parameter $\tau := \tau_n$, we obtain the following result.

Theorem 3.3.2. *Let $\mathcal{X} := [-1,1]^d$, $\beta \in (d/2, +\infty)$, $p \geq 2$, $M > 0$, and $n > 1$, $n \in \mathbb{N}$. Then, for any datasample $\{x_t, y_t\}_{t=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$, any $\varepsilon > 0$ regret of the KAAR with*

$$s = \beta, \quad \tau_n := n^{\frac{d}{2\beta+d}},$$

[4]Throughout the chapter, we refer to constants C, C_1, etc which may depend on the properties of the domain $\mathcal{X}$, the functional class $\mathcal{F}$ or other quantities (such as ε) but are always independent of n. We refer also to $\varepsilon, \varepsilon_1$ as to some infinitesimal numbers (possibly zeros). Their exact values are omitted and may differ from one statement to another, but we will specify this dependency in case this is necessary for analysis.

on the benchmark class $\mathcal{F} := B_{W_p^\beta(\mathcal{X})}(0, R)$ *satisfies the following upper bound*

$$\frac{R_n(\mathcal{F})}{n} \leq Cn^{-\frac{2\beta}{2\beta+d}+\varepsilon} \log(n)\,,$$

where constant C *depends on* $d, s, R, K, M, \mathcal{X}$*, and* ε*, but not on* n*.*

Proof of Theorem 3.3.2 is given in 3.6.5.

Remark 3.3.3. In the lower-bound section we prove that the upper bound of Theorem 3.3.2 matches the minimax optimal for $\beta > d/2$ on the class of bounded Sobolev balls (modulo a constant ε in the exponent that can be made arbitrarily small and logarithmic term in the number of observations). This rate was achieved by Rakhlin and Sridharan (2014) by a non-constructive procedure. An explicit forecaster has been proposed in Gaillard and Gerchinovitz (2015); it can be calculated efficiently when $p = \infty$ and $d = 1$ and in general has exponential time and storage complexity. We believe that Theorem 3.3.2 is the first (essentially) optimal regret upper bound for the classes of bounded balls in Sobolev spaces $W_p^\beta(\mathcal{X})$ with $d \geq 1$, $\beta > \frac{d}{2}$, and $p \geq 2$ that is achieved by a computationally efficient procedure.

3.3.3 Regret upper bound over Sobolev spaces when $\frac{d}{p} < \beta \leq \frac{d}{2}$, $p \geq 2$.

In this part we consider KAAR over the benchmark classes of bounded balls $W_p^\beta(\mathcal{X})$ when $\frac{d}{p} < \beta \leq \frac{d}{2}$, $p > 2$ and refer to this case as a "hard learning" scenario. When $\frac{\beta}{d} \leq \frac{1}{2}$ the Sobolev space $W_p^\beta(\mathcal{X})$ is not (and not embedded into) Sobolev RKHS, so, using KAAR in this case, we must control the error due to using the element $\widehat{f}_t^\tau \in W^s(\mathcal{X})$ when competing against any function from $W_p^\beta(\mathcal{X})$. In this case, the regret analysis can be decomposed into two parts: approximation of any function $f \in W_p^\beta(\mathcal{X})$ by some element $f_\varepsilon \in W^s(\mathcal{X})$ and regret of KAAR with respect to bounded balls in $W^s(\mathcal{X})$. Intuitively, the smaller the approximation error between f and f_ε, the larger the norm of the approximation function f_ε should be, which implies the larger regret upper bound of KAAR with respect to f_ε (see bound (3.11)). Therefore, in this case one has to control a trade-off between the approximation error of $f \in W_p(\mathcal{X})$ by means of some $f_\varepsilon \in W^s(\mathcal{X})$ and the regret suffered of KAAR with respect to any $f_\varepsilon \in W^s$. This is possible and we have the following result.

Theorem 3.3.4. *Let* $\mathcal{X} = [-1, 1]^d$*,* $p > 2$*,* $\beta \in \mathbb{R}_+$*,* $d/p < \beta \leq d/2$*,* $M > 0$*,* $\varepsilon > 0$*,* $n \geq 1$ *and* $\{(x_t, y_t)\}_{t=1}^n \in (\mathcal{X} \times [-M, M])^n$ *be arbitrary sequence of observations. Then by choosing* $s = \frac{d}{2} + \varepsilon$ *and*

$$\tau_n = n^{1-\frac{d\left(1-p^{-1}\right)-\beta'}{d\left(1-2p^{-1}\right)}}$$

where $\beta' = \beta - \varepsilon$ *is sufficiently close to* β *decision rule 1 of KAAR satisfies the following regret upper-bound*

$$R_n(\mathcal{F}) \leq Cn^{1-\frac{\beta}{d}\frac{p-\frac{d}{\beta}}{p-2}+\varepsilon\theta} \log(n),$$

where $\mathcal{F} = B_{W_p^\beta(\mathcal{X})}(0, R)$*, and* $R > 0$*. Constant* C *depends on* $d, s, R, \beta, M, \mathcal{X}$*, and* ε*, but not on* n *and constant* $\theta = \frac{p}{(p-2)d}$*.*

The proof of Theorem 3.3.4 is given in 3.6.5. The theorem and its implications are discussed in Section 3.5. Here we want just to provide two remarks that help to interpret the results.

Remark 3.3.5. In the proof we provide the regret upper bound for any choice $s > \frac{d}{2}$; however the rate for $\frac{d}{p} \leq \beta \leq \frac{d}{2}$ is minimized by the choice $s > \frac{d}{2}$ as small as possible. Therefore, in this situation, we choose $s := d/2 + \varepsilon$ with an arbitrary small $\varepsilon > 0$. Furthermore, in the result of Theorem 3.3.4, the constant C has exponential dependence on the underlying dimension d. To the best of our knowledge, this dependence is unavoidable when using the techniques we use in this work.

Remark 3.3.6. Notice that in an interesting particular case of Theorem 3.3.4 when $p = \infty$ and $\beta \in \mathbb{R}_+$, the space $W_\infty^\beta(\mathcal{X})$ corresponds to functions with derivatives up to order $\lfloor \beta \rfloor$ bounded in supremum norm and $\lfloor \beta \rfloor$-th derivatives are Hölder continuous of order $\alpha \in (0, 1)$ (see (Adams and Fournier, 2003)). Then the regret of Theorem 3.3.4 leads to a regret upper bound of order $O(n^{1-\frac{\beta}{d}+\varepsilon} \log n)$. This upper bound is optimal on the class $W_\infty^\beta(\mathcal{X})$, up to a negligible factor ε that can be chosen arbitrary small (see Section 3.4 for a lower bound on the minimax regret).

3.4 Lower bounds

In this section, we present lower bounds on the regret of any algorithm with respect to any data sequence on the bounded closed balls in Sobolev spaces $W_p^\beta(\mathcal{X})$ with $\beta > d/p$, $p \geq 2$. We define the minimax regret for the problem of online nonparametric regression on the functional class $\mathcal{F}$ as

$$\tilde{R}_n(\mathcal{F}) := \inf_{\mathcal{A}} \sup_{(x_s, y_s)_{s \leq n} \in (\mathcal{X} \times \mathcal{Y})^n} R_n(\mathcal{F}), \tag{3.12}$$

where $\mathcal{A} = (\mathcal{A}_s)_{s \geq 1}$ is any *admissible* forecasting rule, i.e. such that at time $t \in \mathbb{N}$ outputs a prediction $\widehat{y}_t \in \widehat{\mathcal{Y}}$ based on past predictions $(\widehat{y}_s)_{s \leq t-1}$ and data-sample $(\{x_s, y_s\}_{s \leq t-1} \cup x_t)$. More formally, we assume $(\mathcal{A})_{s \geq 1}$ is such that for every $t \in \mathbb{N}$ the map $\mathcal{A}_t : \left(\widehat{\mathcal{Y}}^{t-1} \times (\mathcal{X} \times \mathcal{Y})^{t-1} \times \mathcal{X}\right) \mapsto \widehat{\mathcal{Y}}$ is measurable (with respect to the completion of the product $\sigma-$algebra over the sets $(\mathcal{X} \times \mathcal{Y})^{t-1} \times \mathcal{X}$) and call such algorithm admissible. The most important element of this assumption is that the forecaster cannot use the future outcomes for making decisions at round t. Notice that in this setting we consider the oblivious adversary meaning that all outputs $(x_t, y_t)_{t \geq 1}$ are fixed in advance. With this notation being set we have the following result.

Theorem 3.4.1. *Let $M > 0$, $p \geq 1$ and $\frac{\beta}{d} > 1/p$. Consider the problem of online adversarial nonparameteric regression with $y_t \in [-M, M]$, $x_t \in \mathcal{X} = [-1, 1]^d$ over the benchmark class $\mathcal{F} := B_{W_p^\beta(\mathcal{X})}(0, M)$. Then, minimax regret 3.12 is lower-bounded as*

$$\tilde{R}_n(\mathcal{F}) \geq \begin{cases} C_1 n^{1-\frac{\beta}{d}} & \text{if} \quad \beta \leq \frac{d}{2} \\ C_2 n^{1-\frac{2\beta}{2\beta+d}} & \text{if} \quad \beta > \frac{d}{2} \end{cases},$$

where C_1 and C_2 are constants which depend on $M, \mathcal{X}, d, \beta$, and p, but are independent of n.

The proof is based on the general minimax lower bounds of Rakhlin and Sridharan (2014) and estimation of the fat-shattering dimension of the class $B_{W_p^\beta(\mathcal{X})}(0, M)$ and is given in 3.6.6.

	Statistical i.i.d. regression		Adversarial online nonparametric regression	
	Best known excess risk upper bound	Lower bound	Best known upper bound for $R_n(\mathcal{F})/n$	Lower bound
$\beta > \frac{d}{2}$	$n^{-\frac{2\beta}{2\beta+d}}$	$n^{-\frac{2\beta}{2\beta+d}}$	$n^{-\frac{2\beta}{2\beta+d}}$	$n^{-\frac{2\beta}{2\beta+d}}$
$\frac{d}{p} < \beta \leq \frac{d}{2}$	$n^{-\frac{2\beta}{2\beta+d}}$	$n^{-\frac{1}{2}}$	$n^{-\frac{\beta}{d}}$	$n^{-\frac{\beta}{d}}$
$p = \infty, \beta \leq \frac{d}{2}$	$n^{-\frac{2\beta}{2\beta+d}}$	$n^{-\frac{2\beta}{2\beta+d}}$	$n^{-\frac{\beta}{d}}$	$n^{-\frac{\beta}{d}}$

Table 3.3: Best known regret and excess risk upper and lower bounds on the classes of Sobolev balls. Results achieved by KAAR are highlighted in blue color.

Remark 3.4.2. In Table (3.3) we compare the best known lower and upper regret bounds on the classes of Sobolev balls in the settings of adversarial online regression with the correspondent bounds for the

excess risk in the statistical i.i.d. scenario. Interestingly, on the classes of Sobolev balls in spaces $W_p^\beta(\mathcal{X})$, $\beta \geq \frac{d}{2}$ and Hölder balls $W_\infty^\beta(\mathcal{X})$ rates for the (normalized) regret and for the excess risk are optimal and archived by the regularized empirical risk minimization procedure (for example by regularized least squares estimators in the statistical learning scenario, see Fischer and Steinwart (2017) and KAAR in adversarial regression as shown in this work).

3.5 Discussion

In this part we compare regret rates for KAAR with the existing algorithms in the (adversarial) online nonparametric regression in the terms of regret bounds and computational complexity. Furthermore we compare regret analysis with the excess risk bounds for the known algorithmic schemes in the statistical least-squares regression scenario. We point out on interesting consequences for the gap in the rate which arises through adversarial data.

3.5.1 General comparison to the setting of statistical non-parametric regression

To unify settings we always consider the normalized regret of class $\mathcal{F}$, $\frac{R_n(\mathcal{F})}{n}$. In the statistical setting we assume a sample $\mathcal{D}_n = (z_i = (x_i, y_i))_{i=1}^n$ to be generated independently from the distribution $\nu_{x,y}$ of a pair of random variables X, Y over a probability space $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}(\mathcal{X} \times \mathcal{Y}), \mathbb{P})$ and let $f_{D_n} : \mathcal{X} \mapsto \mathbb{R}$ is some (data-dependent) estimator produced by a measurable learning method $\mathcal{L}$ on $\mathcal{X} \times \mathcal{Y}$. Denote $\nu(y|x)$ to be a regular conditional probability distribution of Y conditional on $\{X = x\}$, and μ to be the $X-$marginal of ν. In the setting of non-parametric regression, for a given class $\mathcal{H} \subset \mathcal{Y}^\mathcal{X}$, the goal is to find a function $f \in \mathcal{H}$ which minimizes the expected squared risk $R_{LS,\nu}(f) = \mathbb{E}_\nu\Big[(Y - f(X))^2\Big]$. The performance measure of the algorithm which outputs decision rule $f_{\mathcal{D}_n}$ in this case is the excess risk, which is $R_{LS,\nu}(f_{\mathcal{D}_n}) - \inf_{f\in\mathcal{F}} R_{LS,\nu}(f)$. If $\mathcal{F}$ is dense in $L_2(\mathcal{X}, \mu)$, it is well-known that the latter is equivalent to minimizing the $\|f_\nu - f_{\mathcal{D}_n}\|_{L_2(\mathcal{X},\mu)}$, where for $\mu-$almost all x, $f_\nu(x)$ is a version of conditional expectation of Y under measure $\nu(\cdot|x)$. Notice that to compare the statistical learning setting and the results of our work we do not necessarily assume that $f_\nu \in \mathcal{H}$. Furthermore,because $f_\nu(\cdot)$ is defined only for $\mu-$ almost all x, we denote f_ν for both the version of this conditional expectation and the correspondent equivalence class with respect to measure μ. We denote $W_p^\beta(\mathcal{X}, \mu)$ the Sobolev space on a probability space $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mu)$. To avoid technical difficulties with threatening weak-derivatives with respect to arbitrary Borel measure, we restrict the space of $X-$marginal probability measures to the subset of all measures which have the Radon-Nikodym derivative with respect to the Lebesgue measure on $\mathcal{X}$. The latter means that the underlying Sobolev space is equivalent to $W_p^\beta(\mathcal{X})$. As before, we consider $\mathcal{X} = [-1, 1]^d$; all the subsequent results in the statistical regression scenario can be reformulated for any bounded subset of $\mathbb{R}^d$ with Lipschitz boundary. We provide comparison to the excess risk upper bounds in high probability, meaning that under $\|f_{\mathcal{D}_n} - f_\nu\|_{L_2(\mathcal{X})} \leq \psi(\delta, n)$ we understand inequality which holds with probability at least $1 - \exp(-\delta)$ for some $\delta > 0$, and $\psi(\cdot, \cdot) : [0, 1] \times \mathbb{R}_+ \mapsto \mathbb{R}_+$ is some function.

We start with the easy problem case, in which $f_\nu \in \mathcal{H}$ and $\mathcal{H}$ is the Sobolev RKHS. Theorem 1 in Caponetto and E.De.Vito (2006) implies (by taking $b = \frac{2\beta}{d}$ and $c = 1$ therein) that for $\mathcal{H} = W^\beta(\mathcal{X})$, $\beta > \frac{d}{2}$, $f_\nu \in W_2^\beta(\mathcal{X}, \mu)$ and $f_\mathcal{D} \in \mathcal{H}$ being a regularized least-squares estimator, we obtain (dropping the constants) that it holds $\|f_{\mathcal{D}_n} - f_\nu\|_{L_2(X,\nu)} \leq Cn^{-\frac{2\beta}{2\beta+d}}$ which is known to be optimal in the setting of non-parametric regression (see Tsybakov (2009) and Győrfi (2002) for matching lower bounds) . Under the same conditions, optimal excess risk rates on $W_2^\beta(\mathcal{X}, \mu)$ can be deduced from Corollary 6 in Lin and Cevher (2018) using the decision rule based on the spectral kernel algorithms or stochastic gradient descent. Notice that in the latter works no assumption on the probability measure ν is posed (a-part the standard in the setting of statistical learning Bernstein condition for $\nu(y|x)$ (see Blanchard and Mücke

(2018), Rudi et al. (2015)) and the variance bound for a random variable Y). It follows that the regret rates of KAAR on classes $W_2^\beta(\mathcal{X})$ match (disregarding the log terms and arbitrary small polynomial factor) the optimal known rates for the excess risk in the i.i.d. scenario on classes $W_2^\beta(\mathcal{X},\mu)$.

A setting in which the underlying RKHS is a subspace of reference class of regular functions is studied in several works. In the particular case of $H := H_\gamma(\mathcal{X})$ being a Gaussian RKHS over $\mathcal{X}$ (which is known to be included in the space of $C_\infty(\mathcal{X})$ functions and is generated by the kernel $k_\gamma(x) = \exp\left(-\frac{\|x\|^2}{\gamma^2}\right)$) and $f_\nu(\cdot) \in W_2^\beta(\mathcal{X}) \cap L_\infty(\mathcal{X})$, $\beta \in \mathbb{R}_+$ Corollary 2 in Eberts and Steinwart (2011) implies that the (Gaussian) kernel ridge regression estimator with the proper choice of *both* regularization parameter λ *and* band-width γ achieves (almost) optimal rates for excess risk of order $n^{-\frac{2\beta}{2\beta+d}+\varepsilon}$ with $\beta > \frac{d}{2}$, $\varepsilon > 0$. The same rates hold when $\beta \leq \frac{d}{2}$ under additional condition $\mathcal{Y} = [-M, M]$ (which implies $\nu-$a.s. boundedness of f_ν that is not ensured unless $\beta > \frac{d}{2}$). In the case $f_\nu \in W_\infty^\beta(\mathcal{X})$ (i.e., its $\lfloor\beta\rfloor-$ derivative is $\beta - \lfloor\beta\rfloor$ Hölder continuous) and $\beta \leq \frac{d}{2}$, excess risk rates in the statistical i.i.d. scenario are optimal (see Chapter 3.2 in Győrfi (2002) for a matching lower bounds and Theorem 14.5 therein). They are better than the normalized regret rates of KAAR in the setting of adversarial regression (which is of order $n^{-\frac{\beta}{d}}$), which are (up to a negligible polynomial factor) optimal (see Theorem 3.4.1 for a matching lower bound). This uncovers an interesting consequence, namely that the gap between the (optimal) rates for regret and excess risk on classes of bounded balls in $W_\infty^\beta(\mathcal{X})$ is due purely to the adversarial nature of the data.

When $f_\nu \in W_2^\beta(\mathcal{X},\mu)$ and the algorithm is the kernelized ridge-regression estimator generated by a kernel of finite smoothness from Corollary 6 in Steinwart (2009) and their discussion afterwards one deduces that excess risk upper bounds of least–squares regression estimator in the Sobolev RKHS $W^s(\mathcal{X})$ with $s \geq \beta > \frac{d}{2}$ are optimal. Notice that in the latter case we do not need to know the smoothness parameter β but only the (possibly crude) upper bound s. Similarly, from Theorem 1 and Example 2 in Pillaud-Vivien et al. (2018), the excess risk rates (in expectation) for the stochastic gradient descent decision rule in Sobolev space $W^s(\mathcal{X})$ on the class $W_2^\beta(\mathcal{X},\mu)$, for $\frac{d}{2} < \beta < s$ can be deduced. They are optimal under the additional assumption $s - \beta \geq \frac{d}{2}$. Corollary 4.4 in **?** implies risk upper bound of order $n^{-\frac{2\zeta}{(2\zeta+\gamma)\vee 1}}$ where parameter ζ is the power of the so-called source condition (see Engl et al. (2000) for more details on source condition and also see Blanchard et al. (2007) for the statistical perspective) and γ is the decay rate of the effective dimension. In the case of Sobolev RKHS $W^s(\mathcal{X})$ and ball in $W_p^\beta(\mathcal{X})$ we have $\zeta = \frac{\beta}{2s}$ and $\gamma = \frac{d}{2s}$ and $\beta \leq \frac{d}{2}$, $s > \frac{d}{2}$) . If $s > d$, we have the excess risk upper bound of order $n^{-\frac{\beta}{s}}$ which is worse than $n^{-\frac{2\beta}{2\beta+d}}$. If $\frac{d}{2} < s \leq d$, the excess risk upper rate is $n^{-\frac{2\beta}{2\beta+d}}$ when $s - \frac{d}{2} < \beta \leq \frac{d}{2}$, and $n^{-\frac{\beta}{s}}$ when $0 < \beta < s - \frac{d}{2}$. In the latter case it is better then the lower bound on the minimax regret but worth then $n^{-\frac{2\beta}{2\beta+d}}$ achieved, as stated above, by, for example, regularized least squares estimator with Gaussian kernels. In the worse case scenario ($\beta < \frac{d}{2}$, $\beta + \frac{d}{2} < s$) one also observe the gap between upper rates for the excess risk in the statistical learning scenario achieved by general spectral regularization methods ($n^{-\beta/s}$) and the lower bounds for the minimax regret ($n^{-\frac{\beta}{d}}$) in the online regression setting. A broader analysis of the difference $f_{\mathcal{D}_n} - f_\nu$ in the norms of the interpolation Hilbert spaces (which can be represented as a range of the fractional power of the kernel integral operator), which range between $\mathcal{H}$ and $L_2(\mathcal{X})$, is provided in Fischer and Steinwart (2017), where the regularized kernel least-squares estimator is considered. Corollary 4.1 therein and inclusion between Sobolev spaces imply excess risk upper bounds of order $n^{-\frac{2\beta}{2\beta+d}+\varepsilon}$ for $f_\nu \in W_p^\beta(\mathcal{X})$, $\beta > 0$, $p \geq 2$. If $\frac{\beta}{d} \in (\frac{1}{p}, \frac{1}{2}]$ and $p \geq 2$, then the aforementioned excess risk rates are better then the regret upper bounds obtained by KAAR (Theorem 3.3.4). To the best of our knowledge, the best–known lower bounds in probability on the excess risk on the classes of balls in the Sobolev spaces are of order $n^{-\frac{1}{2}}$ (see Corollary 4.2 in Fischer and Steinwart (2017) with $t = 0$, $f_\nu \in W_p^\beta(\mathcal{X}) \subset W_2^\beta(\mathcal{X})$ and notice that f_ν is bounded on $\mathcal{X}$ by Sobolev embedding and the Bolzano-Weierstrass theorem).

3.5.2 Comparison in the setting of adversarial nonparametric regression.

Previous works on online nonparametric regression and optimal rates. The setting of online regression when competing against a benchmark of nonparametric functional classes is definitely not new. The standard idea is to use an ε-net of the bounded functional space and exploit the exponential weighted average (EWA) forecaster for a finite class of experts, which will be the element of the $\varepsilon-$net (see Chapter 1 in the monograph Cesa-Bianchi and Lugosi (2006) for the finite EWA and Vovk (2006a) for its application in the case of non-parametric functional classes which are contained in $\mathcal{C}(\mathcal{X})$). Vovk (2006b) analyzes the regret when competing against a general reproducing kernel Hilbert space defined on an arbitrary set $\mathcal{X} \subset \mathbb{R}$ and proves in this case the existence of an algorithm (which is based on the so-called idea of defensive forecasting and requires the knowledge of the feature kernel map) with the regret of order $\mathcal{O}(\sqrt{n})$ over unit balls within the underlying reproducing Hilbert space. Vovk (2007) extends the analysis to the more general framework of Banach spaces, which is described through the decay rate of the so-called modulus of convexity (originally introduced by Clarkson (1936)) and includes, as a particular example, Sobolev spaces with the parameter p of the modulus of convexity being the parameter of the $p-$norm from the definition of $W_\beta^p(\mathcal{X})$. All these approaches have the disadvantage of either having suboptimal regret bounds or having prohibitive computational complexity. Notice that in the framework of online nonparametric regression, *minimax* regret analysis in terms of (sequential) entropy growth rates of the underlying functional classes was provided by Rakhlin and Sridharan (2014). In particular, the optimal rates of order $n^{\frac{d}{2\beta+d}}$ (up to a logarithmic terms) when the reference class is Sobolev RKHS. ($\beta > \frac{d}{2}$) and of order $n^{1-\frac{\beta}{d}}$ on the classes of Hölder balls (which correspond to classes $B_{W_\infty^\beta(\mathcal{X})}(0,R)$) can be achieved by using the generic forecaster with Rademacher complexity as a relaxation (for more details see Example 2, Theorems 2 and 3 and Section 6 in Rakhlin and Sridharan (2014)). Although the relaxation procedure ensures minimax optimality, it is not constructive in general. An explicit forecaster, which designs an algorithm based on a multiscale exponential weighted average algorithm (called Chaining EWA), has been provided in Gaillard and Gerchinovitz (2015) . The latter achieves an optimal rate when competing against functional classes of uniformly bounded functions which have a certain (sharp) growth condition on the sequential entropy (see Rakhlin and Sridharan (2014)). This condition implies optimal rates, for example on classes where sequential entropy is of the order of metric entropy (see 3.6.6 for the definition of the notion of the entropy).

Chaining EWA has been shown to be computationally efficient on the class of Hölder balls ($p = \infty$) with $d = 1$. In general, the Chaining EWA forecaster is computationally prohibitive (as it has exponential time complexity in the number of rounds).

Comparison with Exponential Weighted Average (EWA) forecaster. The idea of using of the EWA forecaster in the nonparametric setting over bounded benchmark functional class $\mathcal{W}$ is to consider the $\varepsilon-$ net $\mathcal{W}_\varepsilon$ of the smallest cardinality:

$$\mathcal{W}_\varepsilon \subset \mathcal{W}, \mathcal{W}_\varepsilon = \min_K \{f_1, f_2, \dots, f_K : \forall f \in \mathcal{W} \exists i \in \{K\}, \text{s.t. } \|f - f_i\|_\infty \leq \varepsilon\}$$

and to use the (finite) exponentially weighted average forecaster (see Cesa-Bianchi and Lugosi (2006)) on the set $\mathcal{W}_\varepsilon$. It was introduced in Vovk (2006a) (see also discussions in Rakhlin and Sridharan (2014) and Gaillard and Gerchinovitz (2015)) and leads to the composed regret upper bound of order $n\varepsilon + \log(\mathcal{N}_\infty(\varepsilon, \mathcal{F}))$, where the last term is the metric entropy of class $\mathcal{F}$ on scale ε. It is known (see Edmunds and Triebel (1996)) that for the benchmark class of Sobolev spaces $W_p^\beta(\mathcal{X})$ (with $p \geq 2$ and $\beta > d/p$), metric entropy is of order $\varepsilon^{-\frac{d}{\beta}}$. Balancing the terms by a proper choice of ε, it results in an upper bound of order $n^{d/(\beta+d)}$ (see also Corollary 8 in Vovk (2006b)). As is illustrated in Figure 3.1 in the $(\frac{\beta}{d}, p^{-1})$ plane, regret upper-bounds of KAAR are smaller than that of EWA as soon as $\frac{\beta}{d}$ is large enough. More precisely, EWA outperforms KAAR when $\frac{\beta}{d} \in [\frac{1}{p}, \frac{\sqrt{1+4p}-1}{2p}]$. The latter is not surprising since KAAR,

which outputs prediction rules in Sobolev RKHS (i.e. functions of sufficiently high regularity), performs worse on the when competing against functions of small regularity. EWA does not have this drawback, as it acts through the space discretization.

In the case $p \geq 2$ and $\frac{\beta}{d} \leq \frac{1}{p}$ it is generally not true that there exists a continuous representative for each equivalence class in $W_p^\beta(\mathcal{X})$. In the case of additional continuity assumption (i.e. considering $W_p^\beta(\mathcal{X}) \cap \mathcal{C}(\mathcal{X})$ as a benchmark class instead) the best (known) upper bound for minimax regret (and thus for regret itself) is of order $n^{1-\frac{1}{p}}$ (see Example 2 in Rakhlin and Sridharan (2014)). It is achieved by a non-constructive algorithm based on the notion of relaxation of sequential Rademacher complexity. Notice that EWA can be also applied over classes of bounded balls in $W_p^\beta(\mathcal{X}) \cap \mathcal{C}(\mathcal{X})$, $\frac{\beta}{d} \leq \frac{1}{p}$; in this case it provides same rate $n^{\frac{d}{\beta+d}}$ which is in this case worth than $n^{1-\frac{1}{p}}$.

Comparison with defensive forecaster by Vovk (2007). Vovk (2007) describes the algorithms that are based on the defensive forecasting schemes in general Banach spaces. The benchmark classes are irregular but continuous functions, particularly including Sobolev spaces. By transferring the results given in Equations (6) and (11) in Vovk (2007) to the setting of this work, defensive forecaster BBK29 (see pages 19–20 in Vovk (2007)) achieves for a unit ball $\mathcal{F} = B_{W_p^\beta(\mathcal{X})}(0,1)$ the following regret bound

$$R_n(\mathcal{F}) \leq \begin{cases} Cn^{1-\frac{\beta}{d}+\varepsilon} & \text{if} \quad p = \infty \\ Cn^{1-\frac{1}{p}} & \text{if} \quad 2 \leq p < \infty \quad \text{and} \quad \frac{d}{p} \leq \beta. \end{cases} .$$

Therefore, in the first case, which corresponds to Hölder balls in $W_\infty^\beta(\mathcal{X})$ and $0 < \beta \leq 1$, we recover the same rate as Theorem 3.3.4 but for the range $\beta > 0$. The rate is optimal, as stated in Theorem 3.4.1. In the second case ($p \geq 2$ and $\frac{d}{p} < \beta < 1$), the upper bounds provided by Theorem 3.3.4 (if $\beta > \frac{d}{p}$) or Theorem 3.3.2 (if $d = 1$ and $\beta > 1/2$) are always better then the correspondent bounds of Vovk (2007).

3.5.3 Computational complexity

Here we consider an optimal computational scheme for KAAR and compare its costs with those of the known nonparametric algorithms (in terms of both runtime and storage complexity).

Recall that KAAR for any $x_t \in \mathcal{X}$, $(x_s, y_s)_{s \leq t-1} \in (\mathcal{X} \times \mathcal{Y})^{t-1}$ computes

$$\widehat{y}_t = \widehat{f}_{\tau,t}(x_t) = \left\langle \widehat{f}_{\tau,t}, k_{x_t} \right\rangle_{\mathcal{H}_k} = \sum_{s=1}^{t} k(x_t, x_s) c_s \,,$$

where $c \in \mathbb{R}^t, c = (K_t + \tau \mathbb{I})^{-1} \tilde{y}_t$, $\tilde{y}_t^\top = \left(Y_{t-1}^\top, 0\right)$ and $K_t = (k(x_i, x_j))_{i,j \leq t}$ is the kernel matrix at step t. A naive way to compute the value of KAAR at the input x_t is by computing the inverse of matrix $K_t + \tau \mathbb{I}_t$. This requires $\mathcal{O}\left(t^3\right)$ iterations in round t and implies $\mathcal{O}\left(n^4\right)$ cumulative time complexity over n rounds. The letter can be improved by using the Cholesky decomposition and the rank-one update of the kernel matrix. Namely, we use the approach as in Algorithm 1 (see Rudi et al. (2015)) for general RKHS. More precisely, at time t we compute the Cholesky decomposition $R_{t-1} R_{t-1}^\top = K_t + \tau \mathbb{I}$; next, we denote the following quantities

$$b_t := (k(x_t, x_1), \ldots, k(x_t, x_{t-1})) \qquad \alpha_t := K_{t-1}^\top b_t + \tau b_t$$
$$\gamma_t := a_t^\top a_t + \tau k(x_t, x_t) \qquad g_t := \sqrt{1 + \gamma_t},$$

and $u_t = (\frac{\alpha_t}{1+g_t}, g_t)$,$v_t = (\frac{\alpha_t}{1+g_t}, -1)$. Using this, we compute an update of R_t:

$$R_t := \begin{pmatrix} R_{t-1} & 0 \\ 0 & 0 \end{pmatrix}, R_t := \text{CHOLUPDATE}(R_t, u_t, \text{'+'}), R_t := \text{CHOLUPDATE}(R_t, v_t, \text{'-'})$$

and calculate the solution's coefficients $c_t = R_t^{-1}\left(R_t^\top\right)^{-1} K_t \tilde{y}_t$. Notice that the procedure CHOLUP $(R, a, " + ")$ returns the upper triangular Cholesky factor of $R + a^\top a$, whereas CHOLUP $(R, a, " - ")$ returns the upper triangule update of $R - a^\top a$. At round t ($t \leq n$) its computational cost is at most $\mathcal{O}(t^2)$. Taking into the account that at the end we compute kernel matrix $K_n = (k(x_i, x_j))_{i,j \leq n}$ for a $d-$dimensional input x_t, which adds dn^2 to the total computational complexity we obtain, that the total computational costs is of the order of $\mathcal{O}(n^3 + n^2 d)$ operations. The latter complexity can be further improved when $\beta > d/(2\sqrt{2} - 2)$ (which implies $\beta > d/2$) to $\mathcal{O}(n^{1+\frac{2d/\beta}{(1-(d/(2\beta))^2)}})$ by using Nyström projection (Jézéquel et al., 2019) while retaining the optimal regret. In particular, it converges to linear runtime complexity when $\beta \to \infty$. Jézéquel et al. (2019) also provides additional improvements to the complexity if features x_t are revealed to the learner beforehand.

As was mentioned before, most existing work in online nonparametric regression on Sobolev spaces (in particular (Rakhlin and Sridharan, 2014; Vovk, 2006a,b, 2007)) does not provide efficient (i.e., polynomial in time) algorithms. Work by Rakhlin and Sridharan (2014) provides an optimal minimax analysis; however, they do not develop constructive procedures. More precisely, they require knowledge of the (tight) upper bounds for the so-called *relaxations*. To obtain the latter ones, in general, one must compute the offset Rademacher complexity, which is numerically infeasible. The approach of using EWA in nonparametric setting (Vovk (2006a)) has non-optimal rates and suffers from prohibitive computational complexity because it must update the weights of the experts in the $\varepsilon-$net. For Sobolev balls its size is of order $\mathcal{O}(\exp(n))$ (given that the number of experts scales as $(\mathcal{N}(\mathcal{F}))$ with $\log \mathcal{N}(\mathcal{F})$ being the metric entropy of the class $\mathcal{F}$, which is polynomial in the number of rounds) so that the total time complexity will be $\mathcal{O}(\exp n + nd)$ (where nd comes from the aggregation of observations $x_t \in \mathcal{X} \subset \mathbb{R}^d$ over n rounds). The defensive forecasting approaches by (Vovk, 2006b, 2007) require the knowledge of the so-called Banach feature map, which is typically inaccessible in the computational design of the algorithm.

To the best of our knowledge, the only algorithm that addresses the problem of computational cost in online nonparametric regression is the Chaining EWA forecaster (Gaillard and Gerchinovitz (2015)). On class $W_\infty^\beta(\mathcal{X})$ with $\beta = r + \alpha$, $\alpha \in (0, 1]$, $r \in \mathbb{N}_*$, the Chaining EWA forecaster can be efficiently implemented through piecewise polynomial approximation —see Lemma 12 and Appendix C in Gaillard and Gerchinovitz (2015). Its time and storage total complexities are of order:

$$\text{Storage:} \quad \mathcal{O}\left(n^{2r+4+\frac{\beta(r-1)+1}{2\beta+1}} \log(n)\right), \qquad \text{Time:} \quad \mathcal{O}\left(n^{(r+1)(2+\frac{\beta}{2\beta+1})} \log(n)\right).$$

Notice that storage complexity of KAAR is $\mathcal{O}\left(n^2\right)$ and it is uniformly better for any $\beta = r + \alpha > 0$ than of Chaining EWA. Furthermore, its time complexity is better for all $\beta \geq 1$ (and worth for $0 < \beta < 1$) than that of the efficient implementation of the Chaining EWA. As was mentioned in Gaillard and Gerchinovitz (2015), in most of the cases the direct implementation of the Chaining EWA forecaster requires $\exp(dpoly(n))$ time (due to the exponentially many updates of the expert's coefficients).

3.6 Proof of the main results of Chapter 3

3.6.1 Approximation properties of the Sobolev spaces.

We recall that $W^s(\mathcal{X})$ is a Sobolev RKHS, a space of continuous representatives from equivalence classes of functions from the Sobolev space $W_2^s(\mathcal{X})$ provided $s > \frac{d}{2}$. The goal of this section is to control the regret with respect to a ball in an arbitrary Sobolev space $W_p^\beta(\mathcal{X})$ with $p \geq 2$ and $\beta \neq s$. To do so, we

need to control the approximation error of $f \in W_p^\beta(\mathcal{X})$ by the elements from some subset $\mathcal{G} \subset W_2^s(\mathcal{X})$ uniformly over $f \in W_p^\beta(\mathcal{X})$. This can be achieved by considering the subset of the band limited functions (see ex. Narcowich et al. (2004)), which is in $W_2^s(\mathcal{X})$ for any $s > 0$. Namely, for $\sigma \in \mathbb{R}_+ \setminus \{0\}$ we define B_σ to be

$$B_\sigma := \{f \in L_2(\mathbb{R}^d) \cap C_\infty(\mathbb{R}^d) : supp(\mathcal{F}(f)) \subset B(0, \sigma)\}, \tag{3.13}$$

where we denote $\mathcal{F}(f)$ for the Fourier transform of f and recall that $B(0, \sigma)$ is an open ball in $\mathbb{R}^d$ with radius σ.

The next result is the consequence of Proposition 3.7 in Narcowich and Ward (2004) (see also the proof of Lemma 3.7 in Narcowich et al. (2004)). To be able to apply the aforementioned Proposition we need to extend functions $f : \mathcal{X} \mapsto \mathbb{R}$, $f \in W^s(\mathcal{X})$ to functions $\tilde{f} : \mathbb{R}^d \mapsto \mathbb{R}$ such that $\tilde{f} \in W^s(\mathbb{R}^d)$. By Stein's Extension Theorem (see Stein (1970), page. 181) because $\mathcal{X}$ is a bounded Lipschitz domain there exists a linear operator $\mathfrak{C} : W^s(\mathcal{X}) \mapsto W^s(\mathbb{R}^d)$ which is continuous (i.e. since it is linear we have $\|\mathfrak{C}f\|_{W_2^s(\mathbb{R}^d)} \leq \tilde{C}\|f\|_{W_2^s(\mathcal{X})}$). For this operator $\mathcal{C}$, every $f \in W^s(\mathcal{X})$ and $g_\sigma \in B_\sigma$ by definition of the norm in $W_2^s(\mathcal{X})$ we have $\|f - g_\sigma\|_{W_2^s(\mathcal{X})} \leq \|\mathfrak{C}f - g_\sigma\|_{W_2^s(\mathbb{R}^d)}$. Applying Lemma 3.7 in Narcowich et al. (2004) to $\mathfrak{C}f \in W^s(\mathbb{R}^d)$, and using the argument as in the proof of Theorem 3.8 in Narcowich et al. (2004) for g_σ given by Lemma 3.7, we have

$$\|f - g_\sigma\|_{W_2^r(\mathbb{R}^d)} \leq c\sigma^{r-s}\|g_\sigma\|_{W_2^s(\mathbb{R}^d)}$$

and

$$\|g_\sigma\|_{W_2^s(\mathcal{X})} \leq \|g_\sigma\|_{W_2^s(\mathbb{R}^d)} \leq c_2\|\mathfrak{C}f\|_{W_2^s(\mathbb{R}^d)} \leq c_3\|f\|_{W_2^s(\mathcal{X})}.$$

Thus we obtain the following statement.

Proposition 3.6.1. *Let $s \geq r \geq 0$. For every $f \in W_2^s(\mathcal{X})$, $\sigma > 0$ there exists a function $g_\sigma \in B_\sigma$ and constants C_0 and C_1 which are independent of σ such that*

$$\|f - g_\sigma\|_{W_2^r(\mathcal{X})} \leq C_0\sigma^{r-s}\|f\|_{W_2^s(\mathcal{X})} \quad \text{and} \quad \|g_\sigma\|_{W_2^r(\mathcal{X})} \leq C_1\sigma^{r-s}\|f\|_{W_2^s(\mathcal{X})}.$$

We now state an upper-bound of $\|f\|_{W_p^r(\mathcal{X})}$ when f belongs to the intermediate Sobolev spaces $W_{p_1}^{s_1}(\mathcal{X})$ and $W_{p_2}^{s_2}(\mathcal{X})$ for some p_1, p_2, s_1, s_2. This result is a Gagliardo-Nirenberg–type inequality and follows from the result originally stated in Theorem 1 in Brezis and Mironescu (2018).

Proposition 3.6.2 (Theorem 1, Brezis and Mironescu (2018))**.** *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a Lipschitz bounded domain. Let $0 \leq r, s_1, s_2 < \infty$ and $1 \leq p_1, p_2, p \leq \infty$ be real numbers such that there exists $\theta \in (0, 1)$ with*

$$r = \theta s_1 + (1 - \theta)s_2 \quad \text{and} \quad \frac{1}{p} = \frac{\theta}{p_1} + \frac{1 - \theta}{p_2}.$$

Let $A := \big\{(s_1, s_2, p_1, p_2) \quad s.t. \quad s_2 \in \mathbb{N}_*,\ p_2 = 1,\ s_2 - s_1 \leq 1 - \frac{1}{p_1}\big\}$. If $(s_1, s_2, p_1, p_2) \notin A$, then there exists a constant $C > 0$ which depends on $s_1, s_2, p_1, p_2, \theta$ and $\mathcal{X}$ such that

$$\|f\|_{W_p^r(\mathcal{X})} \leq C\|f\|^\theta_{W_{p_1}^{s_1}(\mathcal{X})}\|f\|^{1-\theta}_{W_{p_2}^{s_2}(\mathcal{X})},$$

for all $f \in W_{p_1}^{s_1}(\mathcal{X}) \cap W_{p_2}^{s_2}(\mathcal{X})$.

In the next corollary we state two particular cases of Proposition 3.6.2 that will prove useful.

Corollary 3.6.3. *For the domain $\mathcal{X} = [-1, 1]^d$ and any $\varepsilon > 0$, all $p \geq 2$ and $\beta > d/p$ there exists a constant $C > 0$ depending on p, d, ε and β such that*

$$\|g\|_{W_p^{\frac{d}{p}+\varepsilon}(\mathcal{X})} \leq C\|g\|^{\frac{d}{\beta p}+\frac{\varepsilon}{\beta}}_{W_p^\beta(\mathcal{X})}\|g\|^{1-\frac{d}{\beta p}-\frac{\varepsilon}{\beta}}_{L_p(\mathcal{X})}, \tag{3.14}$$

for all function $g \in W_p^\beta(\mathcal{X})$. Furthermore, for all $\beta > 0$, $p \geq 2$ and $\varepsilon > 0$, there exists a constant $C > 0$ depending on β, p, d, and ε such that

$$\|g\|_{W_2^{\frac{d}{2}+\varepsilon}(\mathcal{X})} \leq C \|g\|_{W_2^{\beta p/2}(\mathcal{X})}^{\frac{d+2\varepsilon}{\beta p}} \|g\|_{L_2(\mathcal{X})}^{1-\frac{d+2\varepsilon}{\beta p}}, \tag{3.15}$$

for any function $g \in W_2^\beta(\mathcal{X})$.

Proof First, notice that $\mathcal{X} = [-1,1]^d$ is a Lipschitz bounded domain. The first inequality is obtained by choosing $p_1 = p_2 = p \geq 1$, $r = d/p + \varepsilon$, $s_1 = \beta$, and $s_2 = 0$ in Proposition 3.6.2; checking that $(s_1, s_2, p_1, p_2) \notin A$; and noting that for any $\beta > 0$ we have $W_2^\beta(\mathcal{X}) \cap L_2(\mathcal{X}) = W_2^\beta(\mathcal{X})$. The second inequality stems from the choice $p = p_1 = p_2 = 2$ (note that this is for the p in the Proposition which is different from the p in the inequality), $s_2 = 0$, $s_1 = \frac{\beta p}{2}$ and noting the inclusion $W_2^\beta(\mathcal{X}) \subset W_2^{\beta p}(\mathcal{X}) \subseteq W_2^{\beta p/2}(\mathcal{X}) = L_2(\mathcal{X}) \cap W_2^{\beta p/2}(\mathcal{X})$ which holds true since $p \geq 2$. □

3.6.2 Results from interpolation theory on Sobolev spaces

To provide a sharp upper bound on the effective dimension (Proposition 3.3.1), we also need the following general interpolation result on Sobolev spaces (stated in Theorem 3.8 in Narcowich et al. (2004)). Recall (see Wendlandt (2005), p.172) that the fill distance of a set of points $\mathcal{Z} \subset \mathcal{X}$ is defined as $h_{Z,\mathcal{X}} := \sup_{x \in \mathcal{X}} \inf_{z \in Z} \|x - z\|_2$.

Proposition 3.6.4 (Theorem 3.8 in Narcowich et al. (2004))**.** *Suppose $\Phi : \mathbb{R}^d \to \mathbb{R}$ to be a positive definite function such that its Fourier transform $\mathcal{F}(\Phi)$ satisfies*

$$c_1 \big(1 + \|\omega\|_2^2\big)^{-q} \leq \mathcal{F}(\Phi)(\omega) \leq c_2 \big(1 + \|\omega\|_2^2\big)^{-q} \tag{3.16}$$

where $q \geq s \geq r \geq 0$ and c_1, c_2 are some constants. Assume that $\mathcal{X} \subset \mathbb{R}^d$ is bounded domain, has Lipschitz boundary and satisfies the interior cone condition (see Chapter 4 in Adams and Fournier (2003)) with parameters (φ, R_0). Let $k = \lfloor q \rfloor$ and $\mathcal{Z} \subset \mathcal{X}$ be such that its mesh norm $h := h_{\mathcal{Z},\mathcal{X}}$ satisfies

$$h_{\mathcal{Z},\mathcal{X}} \leq k^{-2} Q(\varphi) R_0, \quad \text{where} \quad Q(\varphi) := \frac{\sin(\varphi)\sin(\theta)}{8(1+\sin(\theta))(1+\sin(\varphi))} \tag{3.17}$$

and $\theta = 2 \arcsin\big(\sin(\varphi)/(4(1+\sin\varphi))\big)$. If $f \in W_2^s(\mathcal{X})$ then there exists a function $v \in \operatorname{span}\{\Phi(\cdot - x_j), x_j \in \mathcal{Z}\}$ such that for every real $0 \leq r \leq s$

$$\|f - v\|_{W_2^r(\mathcal{X})} \leq C h_{\mathcal{Z},\mathcal{X}}^{s-r} \|f\|_{W_2^s(\mathcal{X})}, \tag{3.18}$$

where C is some constant independent of $h_{Z,\mathcal{X}}$ and f.

Let us now instantiate the above Proposition to the specific cases we are interested in by choosing $\mathcal{X}, \Phi, \mathcal{Z}$, and r. Let $T \in \mathbb{N}$ be fixed; set $\mathcal{X} := [-1,1]^d$, Φ being the feature map of Sobolev RKHS $W^s\big(\mathbb{R}^d\big)$. In this case (see 3.1 in Narcowich et al. (2004)) Φ satisfies decay rate from Equation (3.16) with $q = s$. Choose $\mathcal{Z}$ to be the set of points of size T such that $h_{Z,\mathcal{X}} \lesssim T^{-\frac{1}{d}}$ (the latter means that there exists constant $C > 0$ such that $h_{Z,\mathcal{X}} \leq C T^{-\frac{1}{d}}$). To control when then condition 3.17 is fulfilled, we first notice that $\mathcal{X}$ is star-shaped (see Definition 11.25 in Wendlandt (2005), also Proposition 2.1 of Narcowich et al. (2004)); it includes ℓ_2 ball centered at origin with radius $r = 1$ and can be included in the ℓ_2 ball centered at 0 of radius $2\sqrt{d}$. Thus, by Proposition 2.1 in Narcowich et al. (2004), we obtain that $\mathcal{X}$ satisfies interior cone condition with the radius $R_0 = 1$ and angle $\varphi = 2 \arcsin \frac{1}{2\sqrt{d}}$. A

straightforward calculation shows that in this case

$$Q(\varphi) = Q(u(\varphi)) = \frac{u}{8}\left(1 - \frac{8}{8 + u\sqrt{16 - u^2}}\right) = \left(\frac{u}{8}\right)^2 \frac{\sqrt{16 - u^2}}{1 + \frac{u}{8}\sqrt{16 - u^2}},$$

where $u := \frac{\sin\varphi}{1+\sin\varphi} = \frac{\sqrt{4d-1}}{2d+\sqrt{4d-1}}$. Notice that in this case we have that $\frac{1}{8\sqrt{d}} \leq u \leq \frac{1}{2\sqrt{d}}$. We can easily check this by simple inequalities:

$$u = \frac{\sqrt{4d-1}}{2d + \sqrt{4d-1}} \geq \frac{4d-1}{4d} \geq \frac{1}{8\sqrt{d}},$$

and from the other side

$$u \leq \frac{4d-1}{\sqrt{4d-1}} = \frac{1}{\sqrt{4d-1}} \leq \frac{1}{2\sqrt{d}}.$$

From these conditions we deduce $Q(u) \geq \frac{1}{2^{12}d}$. Because $h_{\mathcal{Z},\mathcal{X}} = \sup_{x\in\mathcal{X}} \inf_{z\in\mathcal{Z}} \|x - z\|_2 \lesssim T^{-\frac{1}{d}}$, to satisfy condition (3.17) we need to have $T \geq \left(\frac{k^2}{Q(u)}\right)^d$ where we take $k = \lfloor s \rfloor$ and $R_0 = 1$. Notice that the choice $T \geq \left(4096 s^2 d\right)^d$ ensures the last condition, therefore in order to satisfy condition (3.17) the size T of the grid $\mathcal{Z}$ should be of order $\left(s^2 d\right)^d$. Recall (see Wendlandt (2005)) that the kernel $k(\cdot)$ of the Sobolev space $W^s\left(\mathbb{R}^d\right)$ can be represented by means of Bessel functions of second kind as:

$$k(x_1, x_2) = \frac{2^{1-s}}{\Gamma(s)} \|x_1 - x_2\|_2^{s-\frac{d}{2}} K_{\frac{d}{2}-s}(\|x_1 - x_2\|_2) \tag{3.19}$$

Notice that by Corollary 10.13 in Wendlandt (2005) the norm $\|\cdot\|_{W^s(\mathbb{R}^d)}$ is equivalent to $\|\cdot\|_{W_2^s(\mathbb{R}^d)}$. By Theorem 7.13 in Schaback (2007) (see also Corollary 10.48 on p. 170 in Wendlandt (2005)) a restriction of RKHS $W^s\left(\mathbb{R}^d\right)$ to the domain $\mathcal{X} := [-1, 1]^d$ is itself a RKHS $W^s(\mathcal{X})$ such that it is continuously embedded into $W_2^s(\mathcal{X})$ and its kernel k_1 is a restriction of kernel k to the space $\mathcal{X}$. Thus, we can always consider $W^s(\mathcal{X})$ as a RKHS with reproducing kernel $k_1(\cdot)$ obtained by the restriction of the kernel $k(\cdot)$ given by (3.19) to the domain $\mathcal{X}$. Notice that it can be written as $k_1(x_1, x_2) = \Phi_1(x_1 - x_2)$ and since $\Phi(\cdot)$ satisfies Assumption 3.16 so also $\Phi_1(\cdot)$.

Then, applying Proposition 3.6.4 twice, with $r = 0$ and $r = s$ and the above choices of $\mathcal{X}$, Φ and $\mathcal{Z}$ entails the following corollary.

Corollary 3.6.5. *Let $\mathcal{X} := [-1, 1]^d$, $s > d/2$ and $\mathcal{Z} \subset \mathcal{X}^T$ be a set of points such that fill distance $h_{Z,\mathcal{X}} \lesssim T^{-\frac{1}{d}}$, $T \geq T_0$, $T_0 = \left(4096 s^2 d\right)^d$. Then, for any $f \in W^s(\mathcal{X})$, there exists $\widehat{f} \in \operatorname{span}\{k(x, \cdot), x \in Z\}$, such that*

$$\left\|f - \widehat{f}\right\|_{L_2(\mathcal{X})} \leq C_1 T^{-\frac{s}{d}} \|f\|_{W_2^s(\mathcal{X})}, \qquad \left\|f - \widehat{f}\right\|_{W_2^s(\mathcal{X})} \leq C_2 \|f\|_{W_2^s(\mathcal{X})},$$

and $\widehat{f}(x) = f(x)$ for any $x \in \mathcal{Z}$, where the constants C_1 and C_2 depend on d and s but are independent of the set $\mathcal{Z}$ and function f.

The latter proposition together with Gagliardo-Nierenberg inequality yield the following approximation result of functions $f \in W^s(\mathcal{X})$ by low ranked projections Pf.

Lemma 3.6.6 (Projection approximation)**.** *Let $\mathcal{X} := [-1, 1]^d$, $s > d/2$, $T > T_0$, T_0 is given as in Lemma 3.6.5 and $\mathcal{Z} \subset \mathcal{X}^T$ be a set of points T points $\{x_1, \ldots, x_T\}$ such that the fill distance $h_{\mathcal{Z},\mathcal{X}} \lesssim T^{-\frac{1}{d}}$ and $P_{\mathcal{Z}} : W^s(\mathcal{X}) \to W^s(\mathcal{X})$ be the orthogonal projection on $\operatorname{span}\{k_x : x \in \mathcal{Z}\}$. Then, for any*

$f \in W^s(\mathcal{X})$ and for any $\varepsilon > 0$

$$\|f - P_{\mathcal{Z}} f\|_{L_\infty(\mathcal{X})} = \sup_{x \in \mathcal{X}} |f(x) - (P_{\mathcal{Z}} f)(x)| \leq C T^{-\frac{s-\varepsilon}{d} + \frac{1}{2}} \|f\|_{W^s(\mathcal{X})}, \tag{3.20}$$

where C is a constant independent of f and T. Furthermore, if $s \in \mathbb{N}_*$ then Equation (3.20) holds with $\varepsilon = 0$.

Proof Let $f \in W^s(\mathcal{X})$ and $\varepsilon > 0$. The first inequality follows from inclusion $f - P_{\mathcal{Z}} f \in W^s(\mathcal{X}) \subset C(\mathcal{X})$ when $s > \frac{d}{2}$. Define

$$\widehat{f}_Z := \underset{g \in \text{span}\{k_x, x \in \mathcal{Z}\}}{\text{Arg Min}} \|f - g\|^2_{W^s(\mathcal{X})}. \tag{3.21}$$

Because $W^s(\mathcal{X})$ is a Hilbert space, $\widehat{f}_Z = P_{\mathcal{Z}} f \in W^s(\mathcal{X})$. Furthermore, through reproducing property in RKHS $W^s(\mathcal{X})$ and from the definition of an orthogonal projector, we have for any $x \in \mathcal{Z}$ that $P_{\mathcal{Z}} f(x) = \langle P_{\mathcal{Z}} f, k_x \rangle = \langle f, P_{\mathcal{Z}} k_x \rangle = \langle f, k_x \rangle = f(x)$. By using the Sobolev embedding Theorem between the spaces $W_2^{d/2+\varepsilon}(\mathcal{X})$ and $L_\infty(\mathcal{X})$ (Equation (9) on page 60 in Edmunds and Triebel (1996), applied with $s_1 = d/2 + \varepsilon$, $s_2 = 0$, $n = d$, $p_1 = 2$, and $p_2 = \infty$), and by using Gagliardo-Nierenberg Inequality (3.14), we get

$$\begin{aligned}
\|f - P_{\mathcal{Z}} f\|_{L_\infty(\mathcal{X})} &\leq C_1 \|f - P_{\mathcal{Z}} f\|_{W_2^{\frac{d}{2}+\varepsilon}(\mathcal{X})} \\
&\leq C_2 \|f - P_{\mathcal{Z}} f\|^{\frac{d}{2s}+\frac{\varepsilon}{s}}_{W_2^s(\mathcal{X})} \|f - P_{\mathcal{Z}} f\|^{1-\frac{d}{2s}-\frac{\varepsilon}{s}}_{L_2(\mathcal{X})} \\
&\leq C_3 \big(\|f - P_{\mathcal{Z}} f\|_{W_2^s(\mathcal{X})}\big)^{\frac{d/2+\varepsilon}{s}} T^{-\frac{s}{d}+\frac{1}{2}+\frac{\varepsilon}{d}} \|f\|^{1-\frac{d}{2s}-\frac{\varepsilon}{s}}_{W_2^s(\mathcal{X})} \\
&\leq C_4 T^{-\frac{s}{d}+\frac{1}{2}+\frac{\varepsilon}{d}} \|f\|_{W_2^s(\mathcal{X})}
\end{aligned}$$

where the constants C_1, C_2, C_3, and C_4 are independent of f and T. Finally in the specific case $s \in \mathbb{N}$ we directly apply Corollary 11.33 from Wendlandt (2005) with $m = 0$, $\tau = s$, $q = \infty$ to $f - P_{\mathcal{Z}} f$ and obtain directly bound (3.20) with $\varepsilon = 0$. □

3.6.3 Effective dimension upper-bound for the Sobolev RKHS

Recall that the effective dimension of $\mathcal{H}_k$ based on the data sample $\mathcal{D}$ can be rewritten as

$$d^n_{eff}(\tau) = \text{Tr}\,(T_n + \tau \mathbb{I})^{-1} T_n = \text{Tr}\,(K_n + \tau \mathbb{I})^{-1} K_n,$$

where $T_n = \sum_{s=1}^n k_{x_i} \otimes k_{x_i}$ - (empirical) covariance operator and K_n is the kernel matrix.

Below we provide some auxiliary results that control the tail of the trace of the kernel integral operator. These results are provided in Lemmata 2 and 3 by Pagliana et al. (2020) and are just formulated here for narrative completeness.

Lemma 3.6.7. *Let $\mathcal{H}_k$ be some RKHS over domain $\mathcal{X} \subseteq \mathbb{R}^d$ with continuous reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Let $A : \mathcal{H}_k \to \mathcal{H}_k$ be a bounded linear operator and A^* be its adjoint. Then*

$$\sup_{x \in \mathcal{X}} \|A k_x\|^2_{\mathcal{H}_k} \leq \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \|A^* f\|^2_{L_\infty(\mathcal{X})}.$$

Lemma 3.6.8. *Let $\mathcal{H}_k$ be some RKHS over domain $\mathcal{X} \subseteq \mathbb{R}^d$ with reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and μ be any $\sigma-$finite measure on $\mathcal{X}$. Let $\ell \in \mathbb{N}_+$ and $P : \mathcal{H}_k \mapsto \mathcal{H}_k$ be a projection operator with rank*

less than or equal to $\ell \in \mathbb{N}_+$. Then

$$\sum_{t>\ell} \lambda_t(L) \leq \int_{\mathcal{X}} \|(I-P)k_x\|^2_{\mathcal{H}_k} d\mu(x) \leq \sup_{x\in\mathcal{X}} \|(I-P)k_x\|^2_{\mathcal{H}_k},$$

where $L : L_2(\mathcal{X},\mu) \mapsto L_2(\mathcal{X},\mu)$ is the kernel integral operator as defined in Equation (**??**) *and $\lambda_t(L)$ are its t-th eigenvalues.*

We are now ready to prove our upper bound of the effective dimension. Notice that it can be also recovered from a more general result of Lemma 4 in Pagliana et al. (2020) when taking scale $\gamma = n^{\frac{1}{2s-d}}$ therein. We provide the explicit proof for completeness.

Proof of Theorem 3.3.1. Let $s > d/2$, $t \geq T_0$. By Lemma 3.6.6 for the orthogonal projector P on the set of t points $\mathcal{Z} = \{x_1, \ldots, x_t\} \in \mathcal{X}^t$ such that fill distance $h_{\mathcal{Z},\mathcal{X}} \lesssim t^{-\frac{1}{d}}$ for any $\varepsilon' \in \mathbb{R}_+$ holds

$$\sup_{\|f\|_{W^s(\mathcal{X})}\leq 1} \|f - Pf\|_{L_\infty(\mathcal{X})} \leq C t^{-\frac{s'}{d}+\frac{1}{2}},$$

where $s' = s - \varepsilon'$ and C is a constant that depends on $\mathcal{X}, d, s, \varepsilon$, but not on t. Applying Lemma 3.6.7 with $A = I - P$ we obtain:

$$\sup_{x\in\mathcal{X}} \|(I-P)k_x\|_{\mathcal{H}_k} \leq C t^{-\frac{s'}{d}+\frac{1}{2}}.$$

Let $\{x_i\}_{i\geq 1}^n$ be the sequence of inputs in $\mathcal{X}$. Then, with the choice $\mu := (1/n)\sum_{i=1}^n \delta_{x_i}$, the kernel integral operator L equals K_n/n; combining Lemma 3.6.8 with the last inequality yields

$$\sum_{\ell>t} \lambda_t\big(K_n/n\big) \leq \frac{1}{n}\sum_{i=1}^n \big\|(I-P)k_{x_i}\big\|^2_{\mathcal{H}_k} \leq \sup_{x\in\mathcal{X}} \big\|(I-P)k_x\big\|^2_{\mathcal{H}_k} \leq C t^{-\frac{2s'}{d}+1}. \tag{3.22}$$

From the definition of the effective dimension (see Def. 3.2.1), we have the following upper bound

$$d^n_{eff}(\tau) := \sum_{j=1}^n \frac{\lambda_j(K_n)}{\lambda_j(K_n)+\tau} \leq \sum_{j=1}^t \frac{\lambda_j(K_n)}{\lambda_j(K_n)+\tau} + \tau^{-1}\sum_{j\geq t} \lambda_j(K_n), \tag{3.23}$$

where we used that given that K_n is positive semi-definite, $\lambda_j(K_n) \geq 0$ for all $j \geq 1$. Furthermore, $\lambda_j(K_n)/(\lambda_j(K_n)+\tau) \leq 1$ for all $j \geq 1$, which implies

$$\sum_{j=1}^t \frac{\lambda_j(K_n)}{\lambda_j(K_n)+\tau} \leq t.$$

By homogeneity of the eigenvalues we have $\lambda_j(K_n) = n\lambda_j(K_n/n)$, and therefore

$$\tau^{-1}\sum_{j\geq t} \lambda_j(K_n) = n\tau^{-1}\sum_{j\geq t} \lambda_j(K_n/n).$$

Combining the last two inequalities with Inequalities (3.22) and (3.23), we upper-bound the effective dimension as

$$d^n_{eff}(\tau) \leq t + Cn\tau^{-1}t^{-2s'/d+1}.$$

Choosing t to balance the terms in the above equation (i.e. $t = n^{\frac{d}{2s'}}\tau^{-\frac{d}{2s'}}$), we get

$$d^n_{eff}(\tau) \leq C_1\left(\frac{n}{\tau}\right)^{\frac{d}{2s'}} = C_1\left(\frac{n}{\tau}\right)^{\frac{d}{2(s-\varepsilon')}}.$$

Then, assuming $\varepsilon' < s/2$, and using $1/(1-x) \leq 1 + 2x$ for $0 \leq x \leq 1/2$, we have

$$d_{eff}^n(\tau) \leq C_1 \left(\frac{n}{\tau}\right)^{\frac{d}{2s}\frac{1}{1-\frac{\varepsilon'}{s}}} \leq C_1 \left(\frac{n}{\tau}\right)^{\frac{d}{2s}(1+\frac{2\varepsilon'}{s})} = C_1 \left(\frac{n}{\tau}\right)^{\frac{d}{2s}+\frac{d}{s^2}\varepsilon'} \leq C_1 \left(\frac{n}{\tau}\right)^{\frac{d}{2s}+\frac{2\varepsilon'}{s}}.$$

For any $\varepsilon \in (0,1)$, the choice $\varepsilon' = \varepsilon s/2$ concludes the proof in the case $s \in \mathbb{R}$.

Finally, to satisfy condition $t \geq T_0$ it is sufficient to have n, τ such that $\frac{n}{\tau} \geq C T_0^{\frac{2s'}{d}}$. The latter can be alleviated by additional additive constant in the final bound and is achievable as (when τ_n is chosen as function of n) by an appropriate choice of τ, $\frac{n}{\tau}$ is increasing in n. The result for $s \in \mathbb{R}_+$ follows. Lastly, the result implies also the particular case with $s \in \mathbb{N}$ by taking $\varepsilon = 0$. □

3.6.4 Proof of Theorem 3.3.2

Proof Recall that KAAR, when competing against some function f in an arbitrary RKHS $\mathcal{H}_k$ with a bounded reproducing kernel, attains the general regret upper bound as given in Equation (3.11). Plugging in the bound on the effective dimension of Theorem 3.3.1 with $\mathcal{H}_k = W^s(\mathcal{X})$ into the regret upper bound (3.11) gives

$$R_n(\mathcal{F}) \leq \tau \|f\|_{\mathcal{H}_k}^2 + M^2 C_1 \big(1 + \log\big(1 + \frac{n\kappa^2}{\tau}\big)\big)\big(\widetilde{C}\big(n\tau^{-1}\big)^{\frac{d}{2s}+\varepsilon} + 1\big), \tag{3.24}$$

for any $\varepsilon > 0$. Balancing the first and second terms to minimize the right–hand size (by choosing an appropriate value of τ), (i.e. by setting $\tau := n^{\frac{d}{2s+d}}$), it yields

$$R_n(\mathcal{F}) \leq C n^{\frac{d}{2s+d}+\varepsilon} \log(n),$$

where a constant C depends only on d, s, R, M and $\mathcal{X}$ and does not depend on n. □

3.6.5 Proof of Theorem 3.3.4

We start by introducing a general lemma for the regret of KAAR when competing against continous function and then proceed with the proof of the main theorem.

Lemma 3.6.9. *Let $f \in C(\mathcal{X})$ and $g \in \mathcal{H}_k$. Assume that $(x_i)_{i=1}^n \in \mathcal{X}^n$ and $y_i \in [-M, M]$, for some $M > 0$. Then the regret of algorithm (3.10) when competing against function f is bounded by*

$$R_n(f) \leq \tau \|g\|_{\mathcal{H}_k}^2 + M^2 \left(1 + \log\left(1 + \frac{n\|k\|_\infty^2}{\tau}\right)\right) d_{eff}^n(\tau) + 2n\|f - g\|_{L_\infty(\mathcal{X})}\big(M + \|g\|_{L_\infty(\mathcal{X})}\big)$$

Proof Let $\varepsilon \in (0,1)$ and let $g \in \mathcal{H}_k$ be some function which is to be chosen later. Denote by v the vector $v = (f(x_1), \ldots, f(x_n)) \in \mathbb{R}^n$ and $w = S_n g = (g(x_1), \ldots, g(x_n)) \in \mathbb{R}^n$. We can decompose the regret in the following way:

$$\begin{aligned} R_n(f) &= \big\|Y_n - \widehat{Y}_n\big\|^2 - \|Y_n - v\|^2 \\ &= \big\|Y_n - \widehat{Y}_n\big\|^2 - \|Y_n - w\|^2 - \|v - w\|^2 + 2\langle Y_n - w, v - w\rangle \\ &\leq \big\|Y_n - \widehat{Y}_n\big\|_2 - \|Y_n - w\|^2 + 2\langle Y_n - w, v - w\rangle \\ &= R_n(g) + 2\langle Y_n - w, v - w\rangle. \end{aligned} \tag{3.25}$$

Applying the regret upper bound (3.11) to the element g we get:

$$R_n(g) \leq \tau \|g\|^2_{\mathcal{H}_k} + M^2 \left(1 + \log\left(1 + \frac{n\|k\|^2_\infty}{\tau}\right)\right) d^n_{eff}(\tau),$$

where we recall that $d^n_{eff}(\tau)$ is the effective dimension of the RKHS $\mathcal{H}_k$ with respect to the sample $\mathcal{D} \subset \mathcal{X}^n$. For the second term on the right–hand side in Inequality (3.25) we have:

$$\begin{aligned} \langle Y_n - w, v - w \rangle &\leq \sum_{t=1}^{n} |(y_t - g(x_t))(f(x_t) - g(x_t))| \\ &\leq \sum_{t=1}^{n} (|y_t| + |g(x_t)|)|f(x_t) - g(x_t)| \\ &\leq n\|f - g\|_{L_\infty(\mathcal{X})} \big(M + \|g\|_{L_\infty(\mathcal{X})}\big) \end{aligned} \tag{3.26}$$

Putting together the aforementioned bounds, we obtain our final result. □

Proof of Theorem 3.3.4. Let $\sigma > 0$ be some fixed bandwidth. By Proposition 3.6.1 for any function $f \in W^\beta_p(\mathcal{X}) \subset W^\beta_2(\mathcal{X})$, $p \geq 2$ and $\sigma > 0$ there exists $f_\sigma \in B_\sigma$ such that for $0 \leq r \leq \beta$ we have:

$$\|f - f_\sigma\|_{L_2(\mathcal{X})} \leq C_1 \sigma^{-\beta} \|f\|_{W^\beta_2(\mathcal{X})}, \qquad \|f_\sigma\|_{W^r_2(\mathcal{X})} \leq C_2 \sigma^{(r-\beta)} \|f\|_{W^\beta_2(\mathcal{X})}. \tag{3.27}$$

Because $f \in W^\beta_p(\mathcal{X})$ and $p \geq 2$, so the inclusion implies that we have $\|f\|_{W^\beta_2(\mathcal{X})} \leq C\|f\|_{W^\beta_p(\mathcal{X})}$ with some constant C. Let $\varepsilon_1 > 0$ be any positive number. Applying the Sobolev embedding Theorem (see Equation (9) on page 60 in Edmunds and Triebel (1996) with $s_1 = d/2 + \varepsilon_1$, $s_2 = 0$, $n = d$, $p_1 = 2$, and $p_2 = \infty$), Proposition 3.6.2 for a function $f - f_\sigma \in W^\beta_p(\mathcal{X})$ and the fact that for $p \geq 2$ $W^\beta_p(\mathcal{X}) \subset W^{\beta p/2}_2(\mathcal{X}), W^\beta_p(\mathcal{X}) \subset W^\beta_2(\mathcal{X})$, we get

$$\begin{aligned} \|f - f_\sigma\|_{L_\infty(\mathcal{X})} &\leq C_1 \|f - f_\sigma\|_{W_2^{\frac{d}{2}+\varepsilon_1}(\mathcal{X})} \\ &\leq C_2 \|f - f_\sigma\|^{\frac{d+2\varepsilon_1}{\beta p}}_{W^{\beta p/2}_2(\mathcal{X})} \|f - f_\sigma\|^{1-\frac{d+2\varepsilon_1}{\beta p}}_{L_2(\mathcal{X})} \\ &\leq C_4 \|f - f_\sigma\|^{\frac{d+2\varepsilon_1}{\beta p}}_{W^{\beta p/2}_2(\mathcal{X})} \left(\sigma^{-\beta}\|f\|_{W^\beta_2(\mathcal{X})}\right)^{1-\frac{d+2\varepsilon_1}{\beta p}} \\ &\leq C_5 \|f - f_\sigma\|^{\frac{d+2\varepsilon_1}{\beta p}}_{W^\beta_p(\mathcal{X})} \sigma^{-\beta+\frac{d}{p}+\frac{2\varepsilon_1}{p}} \|f\|^{1-\frac{d+2\varepsilon_1}{\beta p}}_{W^\beta_p(\mathcal{X})}, \end{aligned} \tag{3.28}$$

with a constant C_5, which does not depend on f, f_σ or σ. Because f_σ satisfies (3.27), we obtain for any $r \in \mathbb{R}_+$, $r \geq \beta$:

$$\|f_\sigma\|_{W^r_2(\mathcal{X})} \leq \tilde{C}_1 \sigma^{(r-\beta)} \|f\|_{W^\beta_2(\mathcal{X})} \leq \tilde{C}_2 \sigma^{(r-\beta)} \|f\|_{W^\beta_p(\mathcal{X})}, \tag{3.29}$$

where we obtain the second inequality by inclusion of the Sobolev spaces ($W^\beta_p(\mathcal{X}) \subset W^\beta_2(\mathcal{X})$) and the constant $\tilde{C}_2$ depends only on $\mathcal{X}, d, \beta$ but not σ. Notice that by the triangle inequality and (3.29) with $r = \beta$ we have

$$\|f - f_\sigma\|_{W^\beta_p(\mathcal{X})} \leq \|f\|_{W^\beta_p(\mathcal{X})} + \|f_\sigma\|_{W^\beta_p(\mathcal{X})} \leq \left(1 + \tilde{C}^{\frac{1}{p}}\right) \|f\|_{W^\beta_p(\mathcal{X})}. \tag{3.30}$$

Thus, plugging (3.30) in the Equation (3.28), we deduce:

$$\|f - f_\sigma\|_{L_\infty(\mathcal{X})} \leq C_5 \sigma^{-\beta+\frac{d}{p}+\varepsilon_1} \|f\|_{W_p^\beta(\mathcal{X})}. \tag{3.31}$$

Note also that by using (3.29) with $r = s \geq \beta$ we have:

$$\|f_\sigma\|_{W_2^s(\mathcal{X})} \leq C_2 \sigma^{(s-\beta)} \|f\|_{W_2^\beta(\mathcal{X})} \leq C_3 \sigma^{(s-\beta)} \|f\|_{W_p^\beta(\mathcal{X})}, \tag{3.32}$$

where the last inequality holds because $W_p^\beta(\mathcal{X}) \subset W_2^\beta(\mathcal{X})$. Notice that f_σ, as in Proposition (3.6.1), is of limited bandwidth and is continuous on $\mathcal{X}$; therefore, $\|f_\sigma\|_{L_\infty(\mathcal{X})} = \|f_\sigma\|_{C(\mathcal{X})}$. Now, $f \in W_p^\beta(\mathcal{X})$ and $\beta > \frac{d}{p}$, so by the Sobolev Embedding Theorem $f \in C(\mathcal{X})$; for the f_σ chosen as in Proposition (3.6.1), we have

$$\|f_\sigma\|_{L_\infty(\mathcal{X})} \leq \|f_\sigma\|_{W_p^\beta(\mathcal{X})} \leq \tilde{C}^{1/p} \|f\|_{W_p^\beta(\mathcal{X})},$$

where the last step is true due to (3.29).

Thus, from Lemma 3.6.9 with $g = f_\sigma \in C(\mathcal{X})$, $\mathcal{H}_k = W^s(\mathcal{X})$, we have for the regret of any $f \in W_p^\beta(\mathcal{X})$ it holds that:

$$\begin{aligned} R_n(f) \;\leq\; & \tau \|f_\sigma\|^2_{W_2^s(\mathcal{X})} + M^2 \log\left(e + \frac{en\|k\|^2_\infty}{\tau}\right) d^n_{eff}(\tau) \\ & + 2n\|f - f_\sigma\|_{L^\infty(\mathcal{X})} \big(M + \|f_\sigma\|_{L_\infty(\mathcal{X})}\big). \end{aligned} \tag{3.33}$$

Denote $\varepsilon = \sigma^{-1}$, $s' = s - \varepsilon_1$,$\beta' = \beta - \varepsilon_1$, and plugging (3.31), (3.32), and the bound for $d^n_{eff}(\tau)$ from Theorem 3.3.1 into (3.33) while noticing that $s' - \beta' = s - \beta$, we obtain for any f:

$$\begin{aligned} R_n(f) \;\leq\; & \tilde{C}_1 \, \tau \varepsilon^{-2\left(s'-\beta'\right)} \|f\|^2_{W_p^\beta(\mathcal{X})} + \tilde{C}_2 M^2 \big(1 + \log\big(1 + \frac{n\|k\|^2_\infty}{\tau}\big)\big) n^{\frac{d}{2s'}} \tau^{-\frac{d}{2s'}} \\ & + \tilde{C}_3 n \varepsilon^{\beta'-d/p} \|f\|_{W_p^\beta\left(\mathcal{X}\right)} \left(M + \|f\|_{W_p^\beta(\mathcal{X})}\right) \end{aligned}$$

where $s' = s - \varepsilon_1$,$\beta' = \beta - \varepsilon_1$ and $\tilde{C}_1, \tilde{C}_2, \tilde{C}_3$ are constants depend on d, β, s, d, but not n, M, τ, ε and f. By setting

$$\varepsilon = n^{-\frac{2s'}{2s'(\beta'+d-d/p)-d(\beta'+d/p)}}, \quad \tau = n\varepsilon^{2s'-\beta'-d/p} = n^{1-\frac{2s'(2s'-\beta'-d/p)}{2s'(\beta'+d-d/p)-d(\beta'+d/p)}}$$

and noticing that with such choice of τ, ε for any $f \in \mathcal{F}$ we have $R_n(f) \leq C\tau\varepsilon^{-2\left(s'-\beta'\right)} = n\varepsilon^{\beta'-\frac{d}{p}}$ we obtain for any $f \in \mathcal{F} := \{f \in W_p^\beta(\mathcal{X}) : \|f\|_{W_p^\beta(\mathcal{X})} \leq R\}$

$$R_n(\mathcal{F}) = \sup_{f\in\mathcal{F}} R_n(f) \leq C n^{1-\frac{2s'(\beta'-d/p)}{2s'(\beta'+d-d/p)-d(\beta'+d/p)}} = C n^{1-\frac{\beta' p-d}{\left(\beta' p+d\right)\left(1-\frac{d}{2s'}\right)+d(p-2)}},$$

where C depends on $d, \beta, s, d, R, M, \mathcal{X}$, but not n. Now, to obtain the final claim, we choose $s = \frac{d}{2} + \varepsilon_1$, thus $s' = \frac{d}{2}$, and we have: $1 - \frac{\beta' p-d}{\left(\beta' p+d\right)\left(1-\frac{d}{2s'}\right)+d(p-2)} = 1 - \frac{\beta}{d}\frac{p-\frac{d}{\beta}}{p-2} + \frac{\varepsilon_1 p}{d(p-2)}$, from which the final claim follows. □

3.6.6 Proof of the Theorem 3.4.1

To prove the lower bounds, we use the notion of the *sequential* fat-shattering dimension (see Definition 12 in Rakhlin and Sridharan (2014)). Recall (see Rakhlin et al. (2014)) that a $\mathcal{Z}$-*valued tree* $\mathbf{z}$ *of depth* n is a complete rooted binary tree with nodes labeled by the elements of the set $\mathcal{Z}$. More rigorously, $\mathbf{z}$ is a set of labeling functions $(\mathbf{z}_1, \dots, \mathbf{z}_n)$ such that $\mathbf{z}_t : \{-1,1\}^{t-1} \mapsto \mathcal{Z}$ for every $t \leq n$. For any $\varepsilon \in \{-1,1\}^n$, we denote $\{\mathbf{z}_t(\varepsilon) := \mathbf{z}_t(\varepsilon_1, \dots, \varepsilon_{t-1})\}$ to be the label of the node at the level t, which is obtained by following the path ε.

Definition 3.6.10 (Fat-shattering dimension, see Definition 7 in Rakhlin et al. (2014))**.** Let $\gamma > 0$. An $\mathcal{X}$-valued tree $\mathbf{x}$ of depth d is said to be γ-shattered by $\mathcal{F} = \{f : \mathcal{X} \mapsto \mathbb{R}\}$ if there exists an $\mathbb{R}-$valued tree $\mathbf{s}$ of depth d such that

$$\forall \varepsilon \in \{-1,1\}^d, \quad \exists f^\varepsilon \in \mathcal{F}, \quad \text{s.t.} \quad \varepsilon_t(f^\varepsilon(\mathbf{x}_t(\varepsilon)) - \mathbf{s}_t(\varepsilon)) \geq \frac{\gamma}{2},$$

for all $t \in \{1, \dots, d\}$. The tree $\mathbf{s}$ is called a *witness*. The largest d such that there exists a γ-shattered tree $\mathbf{x}$ is called the (sequential) fat-shattering dimension of $\mathcal{F}$ and is denoted by $\mathrm{fat}_\gamma(\mathcal{F})$.

If the last inequality becomes equality, we say that the tree $\mathbf{x}$ is *exactly* shattered by the elements of $\mathcal{F}$ or (alternatively) that class $\mathcal{F}$ exactly shatters the tree $\mathbf{x}$. We recall also the notion of sequential covering numbers and the sequential entropy of class $\mathcal{F}$.

Definition 3.6.11. A set V of $\mathbb{R}-$valued trees of depth n forms a $\gamma-$ cover (with respect to the ℓ_q norm, $1 \leq q < \infty$) of a function class $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ on a given $\mathcal{X}-$valued tree $\mathbf{x}$ of depth d if

$$\forall f \in \mathcal{F}, \forall \varepsilon \in \{\pm 1\}^d, \exists \mathbf{v} \in V, \text{ s.t. } \left(\frac{1}{n}\sum_{t=1}^{d} |f(\mathbf{x}_t(\varepsilon)) - \mathbf{v}_t(\varepsilon)|^q\right)^{1/q} \leq \gamma.$$

In the case $q = \infty$, we have that $|f(\mathbf{x}_t(\varepsilon)) - \mathbf{v}_t(\varepsilon)| \leq \gamma$ for all $t \in \{1, \dots, d\}$. The size of the smallest γ-cover of a tree $\mathbf{x}$ is denoted by $\mathcal{N}_q(\gamma, \mathcal{F}, \mathbf{x})$; and $\mathcal{N}_q(\gamma, \mathcal{F}, d) = \sup_{\mathbf{x}} \mathcal{N}(\gamma, \mathcal{F}, \mathbf{x})$ where the last supremum is taken over all trees of depth d. Finally, the sequential entropy of class $\mathcal{F}$ is $\sup_{\mathbf{x}} \log \mathcal{N}_q(\gamma, \mathcal{F}, \mathbf{x})$.

To derive the main results of Theorem 3.4.1, we use the following consequences of Lemmata 14,15 in Section 5, Rakhlin and Sridharan (2014).

Lemma 3.6.12 (Variant of **Lemma 14 in Rakhlin and Sridharan (2014)**)**.** *Let $n \in \mathbb{N}_*$, $\mathcal{Y} = [-M, M]$ and $\mathcal{F} \subseteq \{f : \mathcal{X} \to [-M/4, M/4]\}$ for some $M > 0$. If $\gamma > 0$ such that $n \leq \mathrm{fat}_\gamma(\mathcal{F})$ then*

$$\tilde{R}_n(\mathcal{F}) \geq \frac{M}{4} n\gamma.$$

Proof Since, $\gamma > 0$ such that $n \leq \mathrm{fat}_\gamma(\mathcal{F})$, by definition of the fat-shattering dimension there exists an $\mathcal{X}-$valued tree $\mathbf{x}$ of depth n (and a witness of shattering $\boldsymbol{\mu}$), which is shattered by the elements of $\mathcal{F}$. Further proof follows the same lines as in the original argument of Lemma 14 of Rakhlin and Sridharan (2014) with the tree $\mathbf{x}$, witness of shattering $\boldsymbol{\mu}$, $\beta := \gamma$ and functions (as well as witness of shattering bounded in $[-\frac{M}{4}, \frac{M}{4}]$ instead of $[-1, 1]$) therein. □

Lemma 3.6.13 (Variant of **Lemma 15 in Rakhlin and Sridharan (2014)**)**.** *Let $n \in \mathbb{N}_*$, $\gamma > 0$, and $\mathcal{F}'$ be a class of functions from $\mathcal{X}$ to $[-M/4, M/4]$ which exactly γ-shatters some tree $\mathbf{x}$ of depth $\mathrm{fat}_\gamma(\mathcal{F}') < n$. Then the minimax regret with respect to $\mathcal{F}'$ is lower-bounded as*

$$\tilde{R}_n(\mathcal{F}') \geq \frac{M}{4} C\left(2\sqrt{2}\gamma\sqrt{n\mathrm{fat}_\gamma(\mathcal{F}')} - n\gamma^2\right). \tag{3.34}$$

Proof The lemma is proved in the same way as Lemma 15 in Rakhlin and Sridharan (2014), by noting that since $\mathcal{F}'$ exactly shatters $\mathbf{x}$, we can consider $\mathcal{F} = \mathcal{F}'$ in the original proof. The argument follows then the same lines by noticing that the target functional class is a subset of $\{f : \mathcal{X} \mapsto [-\frac{M}{4}, \frac{M}{4}]\}$ (instead of $\{f : \mathcal{X} \mapsto [-1, 1]\}$ as in the original argument). □

To prove the lower bounds, we provide a tight control of $\mathrm{fat}_\gamma(\mathcal{F})$ (in terms of the scale γ, while constants may depend on the range $\mathcal{Y}$,d,β) for $\mathcal{F}$ being the bounded ball in Sobolev space $W_p^\beta(\mathcal{X})$.

We recall the notion of sequential Rademacher complexity (see Rakhlin and Sridharan (2014)):

$$\mathcal{R}_n(\mathcal{F}) = \sup_{\mathbf{x}} \mathbb{E}_\varepsilon \left[n^{-1} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \varepsilon_t f(\mathbf{x}_t(\varepsilon)) \right],$$

where $\mathbb{E}_\varepsilon[\cdot]$ denotes the expectation under the product measure $\mathbb{P} = (\frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1)^{\otimes n}$, the supremum is over all $\mathcal{X}-$ valued trees of depth n. Firstly we provide an auxiliary Lemma which provides an upper bound of the fat-shattering dimension of the Sobolev ball $B_{W_p^\beta(\mathcal{X})}(0, 1)$.

Lemma 3.6.14. *Let $n \in \mathbb{N}, n \geq 1$, $M > 0$ an let $\mathcal{F} := B_{W_p^\beta(\mathcal{X})}(0, M/4)$, $p \geq 2$. For the fat-shattering dimension $\mathrm{fat}_\gamma(\mathcal{F})$ on the scale $\gamma > 0$ when $\beta \neq \frac{d}{2}$ it holds*

$$\mathrm{fat}_\gamma(\mathcal{F}) \leq \max\{\tilde{C}_1 \gamma^{-\left(\frac{d}{\beta} \vee 2\right)}, 1\},$$

where C_1 is some constant which depends on β, d, M but not on γ. In the case $\frac{\beta}{d} = 1/2$ we have

$$\mathrm{fat}_\gamma(\mathcal{F}) \leq \max\{\tilde{C}_2 \left(\frac{\gamma}{\log(\gamma)} \right)^{-2}, 1\},$$

where C_2 is some constant which depends on β, d, M. but not on γ.

Proof of Lemma 3.6.14 Following from Definition (3.6.10), if $\mathbf{x}$ of depth n is $\gamma-$shattered by the elements of $\mathcal{F}$, then $n \leq \mathrm{fat}_\gamma(\mathcal{F})$. For an arbitrary functional class $\mathcal{F}$ from the definition of the fat-shattering dimension for any $\gamma > 0$ such that $\mathrm{fat}_\gamma(\mathcal{F}) > n$ we have that $\mathcal{R}_n(\mathcal{F}) \geq \frac{\gamma}{2}$ (one readily checks this by considering Rademacher complexity over the set of n shattered points). Therefore, $\mathcal{R}_n(\mathcal{F}) \geq \sup\{\frac{\gamma}{2} : \mathrm{fat}_\gamma(\mathcal{F}) > n\}$, which is equivalent to $\mathrm{fat}_\gamma(\mathcal{F}) \leq \min\{n : \mathcal{R}_n(\mathcal{F}) \leq \frac{\gamma}{2}\}$. By Proposition 1 and Definition 3 in Rakhlin et al. (2014) for all $c \in \mathbb{R}$, we have $\mathcal{R}_n(c\mathcal{F}) = |c|\mathcal{R}_n(\mathcal{F})$, where $c\mathcal{F} = \{cf : f \in \mathcal{F}\}$. Taking $c = \frac{4}{M}$, we have for $\mathcal{F}' = B_{W_\infty^\beta(\mathcal{X})}(0, \frac{M}{4})$ that $\mathcal{R}_n(\mathcal{F}') = \frac{M}{4}\mathcal{R}_n(\mathcal{F})$, where $\mathcal{F} = B_{W_\infty^\beta(\mathcal{X})}(0, 1)$. From the definition of $\|\cdot\|_{W_\infty^\beta(\mathcal{X})}$, it follows that if $f \in B_{W_\infty^\beta(\mathcal{X})}(0, 1)$, then $\max_{x \in \mathcal{X}} |f(x)| \leq 1$. By Theorem 3 in Rakhlin et al. (2014) we have for any functional class $\mathcal{F} \subset [-1, 1]^{\mathcal{X}}$

$$\mathcal{R}_n(\mathcal{F}) \leq \sup_{\mathbf{x}} \inf_{\rho \in (0,1]} \left(4\rho + \frac{12}{\sqrt{n}} \int_\rho^1 \sqrt{\log_2 \mathcal{N}_2(\delta, \mathcal{F}, \mathbf{x})} d\delta \right). \tag{3.35}$$

It is straightforward to check that for any tree $\mathbf{z}$ it holds that

$$\mathcal{N}_2(\gamma, \mathcal{F}, \mathbf{z}) \leq \mathcal{N}_\infty(\gamma, \mathcal{F}, \mathbf{z}). \tag{3.36}$$

Furthermore, if $\mathcal{N}_\infty(\mathcal{F}, \gamma)$ is a metric entropy of class $\mathcal{F}$ on scale $\gamma > 0$, then it is easy to check that for any tree $\mathbf{z}$ of depth $d \geq 1$ and any scale $\gamma > 0$, $\mathcal{N}_\infty(\gamma, \mathcal{F}, \mathbf{z}) \leq \mathcal{N}_\infty(\gamma, \mathcal{F})$. Indeed, this follows trivially by taking for any tree $\mathbf{z}$ witness $v(\cdot) = g(\mathbf{z}(\cdot))$, where $g(\cdot)$ is the element of $\gamma-$net such that $\|f - g\|_\infty \leq \gamma$. Furthermore, for $\mathcal{F} = B_{W_p^\beta(\mathcal{X})}(0, 1), \beta > d/p$ the metric entropy of $\mathcal{F}$ on the scale δ is (up to some constant C which does not depend on δ) upper bounded by $\delta^{-\frac{d}{\beta}}$. The latter bound is

a well-known result and it can be deduced from the general result for Besov spaces stated in Theorem 3.5 in Edmunds and Triebel (1996) (see also Equation (38) on page 19 in Vovk (2006b)). Thus, using Equations (3.35) and (3.36), the fact that metric entropy uniformly bounds sequential entropy, properties of Rademacher complexity (see Lemma 3 in Rakhlin et al. (2015)) and the upper bound on the metric entropy of the Sobolev ball $\mathcal{F} = B_{W_\infty^\beta(\mathcal{X})}(0, M/4)$, we get

$$\begin{aligned}\mathcal{R}_n(\mathcal{F}) &= \frac{M}{4}\mathcal{R}_n\left(\frac{4}{M}B_{W_\infty^\beta}\left(0, \frac{M}{4}\right)\right) \leq \frac{M}{4}\mathcal{R}_n\Big(B_{W_\infty^\beta(\mathcal{X})}(0,1)\Big) \\ &\leq \frac{M}{4}\inf_{\rho\in(0,1]}\left(4\rho + \frac{12C}{\sqrt{n}}\int_\rho^1 \delta^{-\frac{d}{2\beta}}d\delta\right) \leq C_1 \inf_{\rho\in(0,1]}\left(4\rho + \frac{12}{\sqrt{n}}\int_\rho^1 \delta^{-\frac{d}{2\beta}}d\delta\right),\end{aligned} \tag{3.37}$$

where we use $C_1 = \frac{M}{4}\max\{1, C\}$ for completeness. Notice that if $\beta > \frac{d}{2}$, then integral $\int_0^1 t^{-\frac{d}{2\beta}}dt$ is finite, thus in this case in (3.37) we can take $\rho = 0$, which implies $\mathcal{R}_n(\mathcal{F}) \leq \frac{12C_1}{\sqrt{n}}\frac{1}{1-\frac{d}{2\beta}}$. When $\beta < \frac{d}{2}$, then the choice $\rho = \rho_{min} = (9n^{-1})^{\frac{\beta}{d}}$ leads to the bound $\mathcal{R}_n(\mathcal{F}) \leq 12C_1 n^{-\frac{\beta}{d}}\frac{1}{1-\frac{2\beta}{d}}$. Finally, in the case when $\beta = \frac{d}{2}$ with the choice $\rho = \frac{3}{\sqrt{n}}$, one gets $\mathcal{R}_n(\mathcal{F}) \leq 6\frac{C_1\ln(n)}{\sqrt{n}}$.

Thus we obtain

$$\mathcal{R}_n\big(\mathcal{F}\big) \leq 12C_1 K n^{-\left(\frac{\beta}{d}\wedge\frac{1}{2}\right)}, \tag{3.38}$$

where in Equation (3.38) $K = \frac{1}{1-(\frac{2\beta}{d}\wedge\frac{d}{2\beta})}$ if $\beta \neq \frac{d}{2}$ otherwise $K = \frac{\ln(n)}{2}$. If $\frac{\beta}{d} \neq \frac{1}{2}$ then we have

$$\begin{aligned}\mathrm{fat}_\gamma(\mathcal{G}) \leq \mathrm{fat}_\gamma(\mathcal{F}) &\leq \min\{n : \mathcal{R}_n(\mathcal{F}) \leq \frac{\gamma}{2}\} \\ &\leq \min\left\{n : 12C_1 K n^{-\left(\frac{\beta}{d}\wedge\frac{1}{2}\right)} \leq \frac{\gamma}{2}\right\} \\ &\leq \left\lceil\left(\frac{\gamma}{24C_1K}\right)^{-\left(\frac{d}{\beta}\vee 2\right)}\right\rceil \\ &\leq \max\{C_2\gamma^{-\left(\frac{d}{\beta}\vee 2\right)}, 1\}.\end{aligned}$$

with $C_2 = 2\cdot(24C_1K)^{\frac{d}{\beta}\vee 2}$. In the case, when $\frac{\beta}{d} = \frac{1}{2}$ we have that by any $n \geq \left\lceil\left(\frac{\gamma/24C_1}{\log(\gamma/24C_1)}\right)^{-2}\right\rceil$ ensures that $\frac{6C_1\ln(n)}{\sqrt{n}} \leq \frac{\gamma}{2}$, from which we deduce $\mathrm{fat}_\gamma(\mathcal{G}) \leq \max\{C_2\left(\frac{\gamma}{\log(\gamma)}\right)^{-2}, 1\}$. □

To derive the first statement of Theorem 3.4.1 we construct a class $\mathcal{G} \subset B_{W_p^\beta(\mathcal{X})}(0, M)$ which satisfies Lemmata 3.6.12 and 3.6.13 and deduce the final bound for the minimax regret $\tilde{R}_n\Big(B_{W_p^\beta(\mathcal{X})}(0, M)\Big)$ by inclusion argument.

Class construction. We provide a class construction, taking inspiration from the nonparametric regression in the statistical learning scenario (see, for example, Theorem 3.2 in Győrfi (2002)). Recall that $\mathcal{X} = [-1, 1]^d$; for a given $n \in \mathbb{N}$ denote $b := n^{-\frac{1}{d}}$. Consider the following set of half–open intervals

$$A = \{A_\ell = [-1 + \ell b, -1 + (\ell+1)b), 0 \leq \ell \leq \lfloor 2n^{1/d}\rfloor - 1\},$$

and let $\mathcal{P} = A^d$ be its $d-$th power. Let $I := \{0, \ldots, \lfloor 2n^{\frac{1}{d}}\rfloor - 1\}^d$, $N = |I| = \lfloor 2n^{\frac{1}{d}}\rfloor^d$ and $\pi : I \mapsto \{1, \ldots, N\}$ be a function which maps an element $k \in I$ to its index in the lexicographic order among the elements in I. Because lexicographic order is a total order, we have that $\pi(\cdot)$ is a bijection. For each $k \in$

$I := \{0, \dots, \lfloor 2n^{\frac{1}{d}} \rfloor - 1\}^d$ such that $\pi(k) = j$, we denote $B_j = \prod_{i=1}^d [-1 + k_i b, -1 + (k_i + 1)b)$. Notice that $\cup_{j=1}^N B_j \subset \mathcal{X}$ and for $i \neq j$ obviously $B_i \cap B_j = \emptyset$. For a cube B_t, $t \in \{1, \dots, N\}$ we denote $a_t \in \mathbb{R}^d$ to be its center. One can show explicitly that $a_t = \big(b\big(\frac{1}{2} + \big(\pi^{-1}(t)\big)_1\big) - 1, \dots, b\big(\frac{1}{2} + \big(\pi^{-1}(t)\big)_d\big) - 1\big)$. Consider the following set of functions:

$$\mathcal{F}_{\beta,d,n} = \left\{ f : f(x) = \frac{M n^{-\frac{\beta}{d}}}{4\|g\|_{W_\infty^\beta(\mathcal{X})}} \sum_{t=1}^N c_t g_{n,t}(x), c_j \in \{-1, 1\} \right\}, \tag{3.39}$$

where $g_{n,t}(x) = g\big(n^{\frac{1}{d}}(x - a_t)\big)$, and g such that $g(x) = \frac{1}{2}\Big(1 - \sigma\big(\frac{\|x\|_2^2 - a^2}{c^2 - a^2}\big)\Big)$, $c = \frac{1}{2}$, $a = \frac{1}{4}$ and $\sigma(t) = \frac{h(t)}{h(t) + h(1-t)}$, $h(t) = e^{-1/t^2}\mathbb{I}_{t>0}$ for $t \in \mathbb{R}$, $x \in \mathbb{R}^d$. We need the following Lemma, which shows that the functional class $\mathcal{F}_{\beta,d,n}$ defined by Equation (3.39) is included in the ball of the space $W_\infty^\beta(\mathcal{X})$.

Lemma 3.6.15. *Let $\beta > 0$, $d \geq 1$, $n \in \mathbb{N}$; consider $N := \lfloor 2n^{\frac{1}{d}} \rfloor^d$ and the class $\mathcal{F}_{\beta,d,n}$, as defined in (3.39). It holds that*

$$\mathcal{F}_{\beta,d,n} \subset B_{L_\infty(\mathcal{X})}\Big(0, \frac{M}{4}\Big).$$

Moreover, a stronger inclusion holds, namely, that

$$\mathcal{F}_{\beta,d,n} \subset B_{W_\infty^\beta(\mathcal{X})}\Big(0, \frac{M}{4}\Big).$$

Proof First, notice that $g(\mathbf{0}) = \frac{1}{2}\big(1 - \sigma(-\frac{a^2}{c^2 - a^2})\big)$. Because $t_0 := -\frac{a^2}{c^2 - a^2} < 0$, $h(t_0) = 0$ and consequently $\sigma(t_0) = 0$ from which we have $g(0) = \frac{1}{2}(1 - \sigma(t_0)) = \frac{1}{2}$. For a cube B_j, if $x \notin B_j$, then we have $g_{n,j}(x) = 0$. Indeed, as $x \notin B_j$, for a_j center of B_j holds $\max_{i \leq d} \Big| x^{(i)} - a_j^{(i)} \Big| \geq \frac{n^{-\frac{1}{d}}}{2}$. Therefore, because $\Big\| n^{\frac{1}{d}} \Big(x^{(i)} - a_j^{(i)} \Big) \Big\|_2 \geq n^{\frac{1}{d}} \max_{i \leq d} |x - a_j| \geq \frac{1}{2}$ and because $g(\cdot)$, as constructed above is a mollifier from $\mathbb{R}^d$ to $\mathbb{R}$ with non-zero support on $B_{\mathbb{R}^d}(0, 1/2)$ (see paragraph 13 in Loring (2011)), we have $g_{n,j}(x) = g\Big(n^{\frac{1}{d}}(x - a_j)\Big) = 0$. From the definition of the norm in the functional class $W_\infty^\beta(\mathcal{X})$, it follows that for any $x \in \mathcal{X}$ we have $|g(x)| \leq \|g\|_{L_\infty(\mathcal{X})} \leq \|g\|_{W_\infty^\beta(\mathcal{X})}$. Furthermore, for any element $f \in \mathcal{F}_{\beta,d,n}$, for any $x \in \mathcal{X} \setminus \cup_{k=1}^N B_k$ we have $f(x) = 0$. If $x \in \cup_{k=1}^N B_k$, then there exists some cube B_j with $x \in B_j$. Thus we get

$$\begin{aligned} |f(x)| &= \left| \frac{M n^{-\frac{\beta}{d}}}{4\|g\|_{W_\infty^\beta(\mathcal{X})}} \sum_{t=1}^N c_j g_{n,t}(x) \right| \leq \frac{M}{4} n^{-\frac{\beta}{d}} \frac{|g_{n,j}(x)|}{\|g\|_{W_\infty^\beta(\mathcal{X})}} \\ &\leq \frac{M}{4} n^{-\frac{\beta}{d}} \frac{\|g\|_{L_\infty(\mathcal{X})}}{\|g\|_{W_\infty^\beta(\mathcal{X})}} \leq \frac{M}{4}, \end{aligned}$$

so that $\mathcal{F}_{\beta,d,n} \subset B_{L_\infty(\mathcal{X})}\big(0, \frac{M}{4}\big)$ and the first part of the claim is proved. Let $\beta = m + \sigma$. For every $r \leq m$, $r \in \mathbb{N}$ and $x \in \mathcal{X}$, we notice that if $x \in \mathcal{X} \setminus \cup_{k=1}^N B_k$, then because it is a finite linear combination of mollifiers we have $D^r f(x) = 0$. By a chain rule for every $f \in \mathcal{F}_{\beta,d,n}$, $x \in \mathcal{X}$, $k \leq N$

such that $x \in B_j$:

$$\begin{aligned}
\sup_{x\in\mathfrak{X}}|D^r f(x)| &= \sup_{B_j\in\mathcal{P}}\sup_{x\in B_j}|D^r f(x)| \\
&= \sup_{B_j\in\mathcal{P}}\sup_{x\in B_j}\frac{M}{4\|g\|_{W_\infty^\beta(\mathfrak{X})}}\left|D^r n^{-\frac{\beta}{d}} g_{n,j}(x)\right| \\
&= \sup_{B_j\in\mathcal{P}}\sup_{x\in B_j}\frac{Mn^{-\frac{\beta}{d}}}{4\|g\|_{W_\infty^\beta(\mathfrak{X})}}\left|D^r g\Big(n^{\frac{1}{d}}(x-a_j)\Big)\right| \\
&= \frac{M}{4\|g\|_{W_\infty^\beta(\mathfrak{X})}} n^{\frac{r-\beta}{d}} \sup_{B_j\in\mathcal{P}}\sup_{x\in B_j}|D^r g(x)| \\
&\le \frac{M}{4}\frac{\sup_{x\in\mathfrak{X}}|D^r g(x)|}{\|g\|_{W_\infty^\beta(\mathfrak{X})}} = \frac{M}{4}\frac{\|D^r g\|_{L_\infty(\mathfrak{X})}}{\|g\|_{W_\infty^\beta(\mathfrak{X})}} \le \frac{M}{4}
\end{aligned}$$

Consider $D^\gamma f$ of a function $f \in \mathcal{F}_{\beta,d,n}$. For some $1 \le j \le N$ we have for any $x, z, \in \overline{B_j}$ (here $\overline{B_j} = B_j \cup \partial B_j$) it holds

$$\begin{aligned}
\frac{|D^\gamma f(x) - D^\gamma f(z)|}{\|x-z\|^\sigma} &= \frac{Mn^{-\frac{\beta}{d}}}{4\|g\|_{W_\infty^\beta(\mathfrak{X})}}\frac{|D^\gamma g_{n,j}(x) - D^\gamma g_{n,j}(z)|}{\|x-z\|^\sigma} \\
&= \frac{Mn^{-\frac{\beta}{d}}}{4\|g\|_{W_\infty^\beta(\mathfrak{X})}}\frac{\left|D^\gamma g\big(n^{\frac{1}{d}}(x-a_j)\big) - D^\gamma g\big(n^{\frac{1}{d}}(z-a_j)\big)\right|}{\|x-z\|^\sigma} \\
&= \frac{Mn^{-\frac{\beta}{d}}}{4\|g\|_{W_\infty^\beta(\mathfrak{X})}}\frac{\left|D^\gamma g(\overline{x})\frac{\partial^\gamma}{\partial x_1\dots\partial x_d}n^{\frac{1}{d}}(x-a_j) - D^\gamma g(\overline{z})\frac{\partial^\gamma}{\partial z_1\dots\partial z_d}n^{\frac{1}{d}}(z-a_j)\right|}{n^{-\frac{\sigma}{d}}\|\overline{x}-\overline{z}\|^\sigma} \\
&\le \frac{Mn^{-\frac{\beta}{d}+\frac{m}{d}+\frac{\sigma}{d}}}{4\|g\|_{W_\infty^\beta(\mathfrak{X})}}\sup_{x,z\in\mathfrak{X},x\neq z}\frac{|D^\gamma g(x) - D^\gamma g(z)|}{\|x-z\|^\sigma} \\
&= \frac{M}{4}\frac{1}{\|g\|_{W_\infty^\beta(\mathfrak{X})}}\sup_{x,z\in\mathfrak{X},x\neq z}\frac{|D^\gamma g(x) - D^\gamma g(z)|}{\|x-z\|^\sigma} \le \frac{M}{4}
\end{aligned}$$

Furthermore, if $B_j, B_k \in \mathcal{P}$ are two different cubes then for $x \in \overline{B_j}$ and $z \in \overline{B_k}$ consider elements $\overline{x} \in \partial B_j$ and $\overline{z} \in \partial B_k$, which lie on the line between x and z. Notice that if $\overline{B_j}$ and $\overline{B_k}$ have common $d-1$ hyperplane (i.e., they are the neighbour cells) then $\overline{x} = \overline{z}$. In all cases, it follows from the

construction of $f \in \mathcal{F}_{\beta,d,n}$ that $D^\gamma f(\overline{x}) = D^\gamma f(\overline{z}) = 0$. Therefore, we have

$$\begin{aligned}
\frac{|D^\gamma f(x) - D^\gamma f(z)|}{\|x-z\|^\sigma} &= \frac{Mn^{-\frac{\beta}{d}}}{4\|g\|_{W^\beta_\infty(\mathcal{X})}} \frac{|D^\gamma g_{n,j}(x) - D^\gamma g_{n,j}(\overline{x}) - D^\gamma g_{n,k}(\overline{z}) + D^\gamma g_{n,k}(z)|}{\|x-z\|^\sigma} \\
&\le \frac{Mn^{-\frac{\beta}{d}}}{4\|g\|_{W^\beta_\infty(\mathcal{X})}} \frac{|D^\gamma g_{n,j}(x) - D^\gamma g_{n,j}(\overline{x})| + |D^\gamma g_{n,k}(\overline{z}) - D^\gamma g_{n,k}(z)|}{\|x-z\|^\sigma} \\
&\le \frac{Mn^{-\frac{\beta}{d}}}{4\|g\|_{W^\beta_\infty(\mathcal{X})}} \frac{\|g\|_{W^\beta_\infty(\mathcal{X})} n^{\frac{\beta}{d}} (\|x-\overline{x}\|^\sigma + \|z-\overline{z}\|^\sigma)}{\|x-z\|^\sigma} \\
&\le \frac{M}{4} 2^\sigma \frac{\|\overline{x}-x\|^\sigma + \|\overline{z}-z\|^\sigma}{2\|x-z\|^\sigma} \\
&\le \frac{M}{4} 2^\sigma \left(\frac{\|x-\overline{x}\| + \|z-\overline{z}\|}{2} \right)^\sigma \frac{1}{\|z-x\|^\sigma} \\
&\le \frac{M}{4} \frac{\|x-z\|^\sigma}{\|x-z\|^\sigma} = \frac{M}{4}
\end{aligned}$$

If for any pair $(x,z) \in \mathcal{X}^2$, $x \neq z$ one element (without losing of generality let it be z) does not belong to the union of the cubes $\cup_{B\in\mathcal{P}} B$, then we can substitute this point by the point $\overline{z}$, which is the intersection of the segment $[x,z]$ and the boundary of the closest cube to the point z. Notice that in this case $D^\gamma f(z) = D^\gamma f(\overline{z}) = 0$ by construction of f and $\|x-z\|_2^\sigma \ge \|x-\overline{z}\|_2^\sigma$. Applying aforementioned analysis to a pair $(x,\overline{z})$ which lies in some (different) cubes B_j, B_k, we get

$$\frac{|D^\gamma f(x) - D^\gamma f(z)|}{\|x-z\|^\sigma} \le \frac{|D^\gamma f(x) - D^\gamma f(\overline{z})|}{\|x-\overline{z}\|^\sigma} \le \frac{M}{4}.$$

Finally, case $(x,z) \in \mathcal{X}^2$, where none of the points belong to the union of the cubes, is trivial.

Considering these cases together we have $\sup_{x,y\in\mathcal{X}, x\neq y} \frac{|D^\gamma f(x) - D^\gamma f(y)|}{\|x-y\|^\sigma} \le \frac{M}{4}$ for any $f \in \mathcal{F}_{\beta,d,n}$. Therefore, $\mathcal{F}_{\beta,d,n} \subset B_{W^\beta_\infty(\mathcal{X})}\big(0, \frac{M}{4}\big)$.

□

Proof of Theorem 3.4.1. For $n \ge 1$ consider the functional class $\mathcal{F}_{\beta,d,n}$ as given by Equation 3.39. Consider a $\mathcal{X}-$valued tree $\mathbf{x}$ of depth $N := \lfloor 2n^{\frac{1}{d}} \rfloor^d$ constructed as follows: for any $\varepsilon \in \{-1,1\}^N$, any $t \le N$ we set $x_t(\varepsilon) = a_t$, where a_t is the center of the correspondent cube. Now, for any $\varepsilon \in \{-1,1\}^n$ consider $f^\varepsilon(\cdot) \in \mathcal{F}_{\beta,d,n}$ where $\mathcal{F}_{\beta,d,n}$ as in (3.39) and $f^\varepsilon(x) = \frac{M}{4\|g\|_{W^\beta_\infty(\mathcal{X})}} \sum_{j=1}^N \varepsilon_j n^{-\frac{\beta}{d}} g_{n,j}(x)$. Then for the tree $\mathbf{x}$, for every $\varepsilon \in \{-1,1\}^N$, $1 \le t \le N$ and a real-valued (witness of shattering) $s_t(\cdot) := 0$, we have

$$\begin{aligned}
\varepsilon_t(f^\varepsilon(x_t(\varepsilon)) - s_t(\varepsilon)) = \varepsilon_t f^\varepsilon(x_t(\varepsilon)) = \varepsilon_t f^\varepsilon(a_t) &= \frac{M}{4\|g\|_{W^\beta_\infty(\mathcal{X})}} \varepsilon_t \sum_{j=1}^N \varepsilon_j n^{-\frac{\beta}{d}} g_{n,j}(a_t) \\
&= \frac{M}{4\|g\|_{W^\beta_\infty(\mathcal{X})}} n^{-\frac{\beta}{d}} g_{n,t}(a_t) \\
&= \frac{M}{4\|g\|_{W^\beta_\infty(\mathcal{X})}} n^{-\frac{\beta}{d}} g(0) = C_{M,g} \frac{n^{-\frac{\beta}{d}}}{2},
\end{aligned} \tag{3.40}$$

where $C_{M,g} := \frac{M}{4\|g\|_{W^\beta_\infty(\mathcal{X})}}$. Thus, class $\mathcal{F}_{\beta,d,n}$ with $\tilde{\gamma} = \tilde{\gamma}(n) := C_{M,g} n^{-\frac{\beta}{d}}$ (exactly) shatters the tree

$\mathbf{x}$. Notice that $N = \lfloor 2n^{\frac{1}{d}} \rfloor^d \leq 2^d n$; from the other side we have $N \geq \left(n^{\frac{1}{d}}\right)^d \geq n$. Thus, from the definition of fat-shattering dimension, it follows,

$$\mathrm{fat}_{\tilde{\gamma}}\left(\mathcal{F}_{\beta,d,n}\right) \geq N \geq n. \tag{3.41}$$

All conditions of Lemma 3.6.12 are fulfilled for the class $\mathcal{F}_{\beta,d,n}$; by Lemma 3.6.15 $\mathcal{F}_{\beta,d,n} \subset B_{W^\beta_\infty(\mathfrak{X})}\left(0, \frac{M}{4}\right)$. Applying Lemma 3.6.12 to the class $\mathcal{F}_{\beta,d,n}$, using Lemma 3.6.15 and simple inclusion $B_{W^\beta_\infty(\mathfrak{X})}\left(0, \frac{M}{4}\right) \subset B_{W^\beta_p(\mathfrak{X})}\left(0, \frac{M}{4}\right) \subset B_{W^\beta_p(\mathfrak{X})}(0, M)$, we obtain for the Sobolev ball $\mathcal{F} := B_{W^\beta_p(\mathfrak{X})}(0, M)$

$$\tilde{R}_n(\mathcal{F}) \geq \tilde{R}_n(\mathcal{F}_{\beta,d,n}) \geq \frac{M}{4} n\tilde{\gamma} \geq \frac{\cdot M^2}{16\|g\|_{W^\beta_\infty(\mathfrak{X})}} n^{1-\frac{\beta}{d}},$$

so that the case $\frac{d}{p} < \beta \leq \frac{d}{2}$ is proved.

To prove the second bound, notice that by Lemma (3.6.15) for any $n \in \mathbb{N}_*$, $\mathcal{F}_{\beta,d,n} \subset B_{W^\beta_\infty(\mathfrak{X})}\left(0, \frac{M}{4}\right)$, which implies that $\mathrm{fat}_\gamma\left(B_{W^\beta_\infty}\left(0, \frac{M}{4}\right)\right) \geq \mathrm{fat}_\gamma(\mathcal{F}_{\beta,d,n})$. In particular, this holds if we choose $n_0 := \left\lfloor \left(\frac{\gamma}{C_{M,g}}\right)^{-\frac{d}{\beta}} \vee 1 \right\rfloor$, then $n_0 < \left(\frac{\gamma}{C_{M,g}}\right)^{-\frac{d}{\beta}} \vee 1$, which is equivalent to $C_{M,g} n_0^{-\frac{\beta}{d}} \leq \gamma$. Notice that if $\gamma_1 < \gamma_2$ then $\mathrm{fat}_{\gamma_1}(\mathcal{F}) \geq \mathrm{fat}_{\gamma_2}(\mathcal{F})$. Applying the first property to the classes $\left(B_{W^\beta_\infty}\left(0, \frac{M}{4}\right)\right)$ and $\mathcal{F}_{\beta,d,n_0}$ on the scale γ and the second property for the class $\mathcal{F}_{\beta,d,n_o}$ on the scales γ and $C_{M,g} n_0^{-\frac{\beta}{d}}$ we consequently get

$$\mathrm{fat}_\gamma\left(B_{W^\beta_\infty(\mathfrak{X})}\left(0, \frac{M}{4}\right)\right) \geq \mathrm{fat}_\gamma(\mathcal{F}_{\beta,d,n_0}) \geq \mathrm{fat}_{C_{M,g} n_0^{-\frac{\beta}{d}}}(\mathcal{F}_{\beta,d,n_0}) \geq n_0. \tag{3.42}$$

Finally, because $n_0 \geq 1$, so by using elementary $\lfloor a \rfloor \geq \frac{a}{2}$ we have $n_0 \geq \frac{1}{2}\left(\frac{\gamma}{C_{M,g}}^{-\frac{d}{\beta}} \vee 1\right)$; therefore,

$$\mathrm{fat}_\gamma\left(B_{W^\beta_\infty}(0, M/4)\right) \geq \mathrm{fat}_\gamma(\mathcal{F}_{\beta,d,n_0}) \geq \frac{1}{2}\left(\left(\frac{\gamma}{C_{M,g}}\right)^{-\frac{d}{\beta}} \vee 1\right)$$

.

Choose $\gamma := C_{M,g}^{\frac{d}{\beta+d}} n^{-\frac{\beta}{2\beta+d}}$, $n_0 := \left\lfloor \left(\frac{\gamma}{C_{M,g}}\right)^{-\frac{d}{\beta}} \right\rfloor$, where C_1 is a constant as in Lemma 3.6.14 and $C_{M,g}$ is a constant as in Equation (3.40). For $\beta > \frac{d}{2}$, we have by inclusion and by Lemma 3.6.14 that for any n with the choice of γ as before it holds: $\mathrm{fat}_\gamma(\mathcal{F}_{\beta,d,n}) \leq \mathrm{fat}_\gamma\left(B_{W^\beta_\infty(\mathfrak{X})}\left(0, \frac{M}{4}\right)\right) \leq \tilde{C} n^{\frac{2\beta}{2\beta+d}} < \tilde{C} n$. Furthermore, as for any $n \in \mathbb{N}$, $B_{W^\beta_\infty(\mathfrak{X})}\left(0, \frac{M}{4}\right) \supset \mathcal{F}_{\beta,d,n}$, so, in particular, $B_{W^\beta_\infty(\mathfrak{X})}\left(0, \frac{M}{4}\right) \supset F_{\beta,d,n_0}$, which implies $R_n\left(B_{W^\beta_\infty(\mathfrak{X})}\left(0, \frac{M}{4}\right)\right) \geq R_n\left(\mathcal{F}_{\beta,d,n_0}\right)$. Thus, applying Lemma 3.6.13 to the class $\mathcal{F}_{\beta,d,n_0}$ with any n and γ, n_0 as above, we obtain:

$$\begin{aligned}
\tilde{R}_n(\mathcal{F}_{\beta,d,n_0}) &\geq C\gamma\sqrt{n}\left(2\sqrt{2}\sqrt{\mathrm{fat}_\gamma(\mathcal{F}_{\beta,d,n_0})} - \sqrt{n}\gamma\right) \\
&\geq C n^{-\frac{\beta}{2\beta+d}} \sqrt{n}\left(2C_{M,g}^{\frac{d}{d+\beta}} n^{\frac{d}{2(2\beta+d)}} - C_{M,g}^{\frac{d}{\beta+d}} n^{\frac{1}{2}-\frac{\beta}{2\beta+d}}\right) \\
&\geq \tilde{C}_1 n^{-\frac{\beta}{2\beta+d}+\frac{1}{2}+\frac{d}{2(2\beta+d)}} = \tilde{C}_1 n^{\frac{d}{2\beta+d}},
\end{aligned}$$

where C_1 is some constant independent of n. Now, the final bound for $\beta < \frac{d}{2}$ follows from the inclusion $\mathcal{F}_{\beta,d,n_0} \subset B_{W_\infty^\beta}\left(0, \frac{M}{4}\right) \subset B_{W_p^\beta(\mathcal{X})}\left(0, \frac{M}{4}\right)$ which implies $\tilde{R}_n\Big(B_{W_\infty^\beta(\mathcal{X})}(0, M)\Big) \geq \tilde{R}_n\big(B_{W_p^\beta(\mathcal{X})}(0, M)\big) \geq \tilde{R}_n\big(\mathcal{F}_{\beta,d,n_0}\big)$. □

3.6.7 Regret rates comparison

Here we provide a short comparison of the exponents of theoretical regret rates between KAAR (3.10) and EWA (Vovk (2006a)). One can check that when $\frac{\beta}{d} < \frac{\sqrt{1+4p}-1}{2p}$, EWA provides better rate than KAAR, given by (3.10) with $s = \frac{d}{2} + \varepsilon$, $\varepsilon > 0$ and τ_n chosen as in the Theorem 3.3.4. For a fixed pair (β, d) this means that with increasing regularity of the function f in terms of its integral $p-$norm, KAAR estimates its behaviour better than EWA for a larger range of possible values (β, d). This effect is illustrated in Figure 3.2.

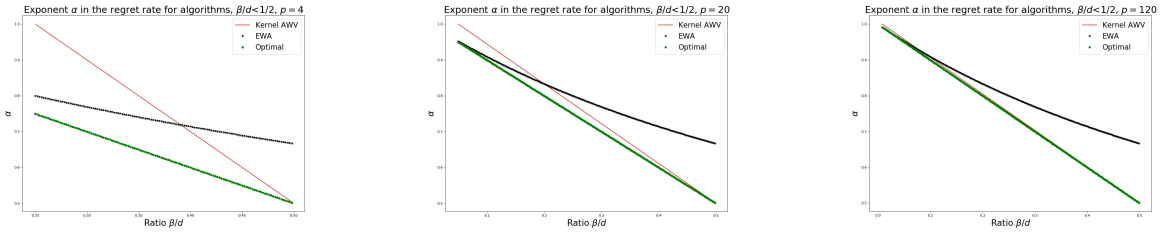

Figure 3.2: Exponent of the regret in the case $W_p^\beta(\mathcal{X})$, $\frac{1}{p} < \frac{\beta}{d} \leq \frac{1}{2}$, $p = 4, 20, 120$.

Chapter 4

Restless stationary bandits with dependencies

Contents

This chapter is devoted to the study of the multi-armed bandit problem (see Chapter 1) in a case in which the arm samples were dependent over time and generated from a $\mathcal{C}-$mixing process. In particular, over the set of all index-based switching arm strategies (i.e. those which concentrate on the choosing of the best arm with respect to some score function which is updated during the course of the game) we consider the improved UCB algorithm (see Auer and Ortner (2010) also see Perchet and Rigollet (2013)) which is based on the sequential arm-elimination policy (see Evan-Dar et al. (2006) for the first reference). We analyse the regret of this policy (both problem-dependent and problem-independent) in the settings of different decay rates of mixing coefficients. Materials of this chapter are based on the joint work with Gilles Blanchard and Alexandra Carpentier, see Zadorozhnyi et al. (2019).

4.1 Introduction

Recall from the introduction that a multi-armed stochastic bandit problem can be modelled as a sequential game between learner and environment which evolves as follows. At every round $t \leq T$ the learner, based on the decision $I_t = a$, selects an action ("pulls an arm") $a \in \{1, \ldots, K\}$ and observes a (stochastic) version of the payoff of round t based on the observation X_t^a from a stochastic process $(X_t^a)_{t\geq 1}$. The general goal of this game is to find a sequence of pulling arms $(I_t)_{t\geq 1}$ that maximizes the (cumulative) payoff over some time horizon T, $\sum_{t=1}^{T} X_t^{I_t}$ based on the observed data and decisions. The performance measure of the strategy is typically (but, as we will discuss later, not exclusively) restricted to the notion of cumulative regret with respect to pulling of the best action in hindsight. Its formal definition is recalled and discussed in Section 4.2. In this chapter we assume that the payoffs are generated from the trajectory of some stochastic process and thus consider the setting of stochastic bandits. Notice that numerous studies (see for example Auer (2002), Bubeck (2010), Auer et al. (2002) and Gerchinowitz and Lattimore (2016)) have been conducted in the so-called setting of adversarial bandits, where no stochastic assumption on the sequences $(X_t^a)_{t\geq 1}$ has been made. Since its emergence (Robbins (1952)) in the literature, in the setting of stochastic multi-armed bandit problems, it has been commonly assumed that the outcomes of the arms are stochastically independent. This means that for each round t, the distribution of X_t^k (outcome of arm k at moment t) is stationary and does not depend on the history $\{X_s^k, s < t\}$ of the previous outcomes. From the application perspective, however, many of the real-world problems in which bandits find their use display an intrinsic dependence between the sequence of future outcomes and past realizations. For example, in the ad-placement problem, whether a user clicks on a given ad in the near future depends closely on whether that user has clicked on the ad at the current time. If the user becomes bored, he will not be willing to choose this ad again immediately; however, as time passes, the user is more likely to click on the ad once more. In such a scenario, the correlation between previous and future observations decays as the time gap increases. Additionally, in this example, the decision to ultimately make a purchase might not be affected by the present minute fluctuations of the user's interest, instead depending only on its *nominal average* level, for which the long-term averaged clicking rate is a proxy. Another motivating example is the cognitive radio problem (see also Ortner et al. (2014) for the same motivation) in which each arm of the bandit instance can be seen as a radio channel which at each time-point can be seen as busy or occupied. At each time point a learner can choose only one channel, see the state of only one channel and (if the channel is not occupied) send the message via it. It is natural to assume that whether the channel is available in the moment depends on the past. Furthermore, the dynamics of all of the processes depends on time and it evolves even if the arm is not pulled. This can be modelled through the assumption on the decaying correlations between the present and past of stochastic processes, and through the assumption of restless sample generation on the bandit instance.

Generalization of the stochastic bandits to the setting with dependent outcomes was first considered by Whittle (1988). If the underlying stochastic processes are Markov chains (which, as we will show later, satisfy the weak-dependency assumption 4.4) with known dynamics, the regret has been studied by Guha et al. (2010b) and Ortner et al. (2014). Problem-dependent asymptotic pseudo-regret upper bounds for the rewards generated from so-called $\varphi-$mixing processes have been derived by Audiffren and Ralaivola (2015). In the latter study, authors have devised a UCB-type strategy and considered scenarios of both fast and slowly mixing arms (this notion will be introduced later and depend on the correlation decay rate between past and future of the corresponding process). Under the same weak dependency assumption of φ–mixing, the work of Grünewälder and Khaleghi (2017) has extended the analysis to the different regret concepts (namely for when the possible policies are switching arm policies), providing an upper bound analysis in the so-called fast φ-mixing setting.

Overview of the main results. In this chapter we considered the general notion of weakly dependent processes under a $\mathcal{C}-$mixing assumption and a version of a known IMPROVED-UCB ALGORITHM (see

Auer and Ortner (2010) and also Evan-Dar et al. (2006) for a general introduction to the concept of arm-elimination strategies). To give a short overview of the contribution of the chapter we note that, in the fast mixing scenario, the obtained upper bounds match (up to a multiplicative constant) the independent case. The contamination term due to dependency in the regret upper bound comes *additively* and does not scale with the number of arms. For the problem-dependent upper bound, it only depends on the choice of the threshold error level in the bound and on the mixing rate. Namely, in the problem-dependent case we have that for the regret (see Chapter 1 for the definition) over a $\mathcal{C}-$ mixing bandit instance $\mathbb{P}$ of $\mathcal{C}-$Mix UCB the following problem-dependent bound

$$R_{\mathbb{P}}(T) \leq M \sum_{k=1}^{K} \Big(\Delta_k + \frac{96}{\Delta_k} + \frac{32 \log\big(T\Delta_k^2\big)}{\Delta_k} \Big) + 64 \sum_{k \in A_0 \setminus A_\lambda} \frac{1}{\lambda} + T \max_{k \in A_0 \setminus A_\lambda} \Delta_k,$$

and the following problem-independent bound

$$R_{\mathbb{P}}(T) \leq C \sqrt{KT \log(K)},$$

where $\lambda > \frac{1}{\sqrt{T}}$ is some threshold, $A_\lambda = \{k : \Delta_k \geq \lambda\}$ and M, C are some constants which depend on the mixing rate but are independent of T, K. Furthermore, in the so-called slowly mixing regime (when $\phi_{\mathcal{C}}(t) \sim t^{-\alpha}$, $\alpha < \frac{1}{2}$) we get

$$R_{\mathbb{P}}(T) \leq C_1 \sum_{k \in A_\lambda} \max\{\frac{\log\big(\mathcal{A}T\Delta_k^2\big)}{\Delta_k}, 1\} + C_2 (\Delta_{*,\lambda})^{1-\frac{1}{\alpha}} \big(C_3 \log\big(\mathcal{A}T\Delta_{*,\lambda}^2\big)\big)^{\frac{1}{2\alpha}} + \frac{12}{\sqrt{e}} \sum_{k \in A_0 \setminus A_\lambda} \frac{1}{\lambda} + T \max_{k : A_0 \setminus A_\lambda} \Delta_k,$$

and for the problem-independent case

$$R_{\mathbb{P}}(T) \leq C_0 \sqrt{T} \max\{\sqrt{K \log(T)}, T^{1/2-\alpha} (\log(T))^{\frac{1}{2\alpha}}\},$$

where $\Delta_{*,\lambda} = \min_{k \in A_\lambda} \Delta_k$ and C_0, C_1, C_2, C_3 are some constants whose values depend on α but not on T or K. In the problem-dependent regret upper bounds for the slow mixing scenario the main regret term (similar in order to the i.i.d. case) comprises a sum over arms of the inverse arm's gaps; the contamination term due to dependency remains negligiblein the regime when there is a large number of suboptimal arms whose expected payoff is close to the chosen threshold level. This can be intuitively understood, given that the time between two pulls of the same arm will typically remain larger than the correlation distance in that situation. For the problem-independent upper bound, the additive penalty resulting from dependency is determined by the relation among the number of arms, the exponent of the polynomially mixing process, and the time-horizon. In the slow mixing scenario, in the case of problem-dependent and problem independent bounds, it allows us to derive bounds which (in certain regimes) match their independent data analogues for the pseudo-regret.

In Section 4.2, we recall the concept of a weak-mixing process (see Chapter 1) and introduce the corresponding probabilistic toolbox for the real-valued discrete processes. In Section 4.3, we present the $\mathcal{C}-$MIX UCB learning algorithm and provide the problem–dependent and problem–independent bounds on its regret. Depending on the correlation decay rates we distinguish between scenarios of *slow-* and *fast-* mixing. In Section 4, we discuss the optimality of the given bounds in the *fast mixing scenario* and derive the lower bounds (which are also optimal up to a logarithmic factor of the number of rounds) for the case of slowly-mixing processes. Finally, in Section 4.5 we discuss the obtained results and show their advantages over the existing bounds, pointing out the cases of slow mixing processes in which i.i.d. regret upper bounds can be recovered. All proofs are postponed to the end and are given in Section 4.7.

4.2 Setting and preliminaries

Let $\mathcal{X} = [0,1]$ and $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ – be a measurable space equipped with a Borel $\sigma-$algebra $\mathcal{B}(\mathcal{X})$. Denote $K, T \in \mathbb{N}$ as the number of arms and the number of rounds (time-horizon). We always assume that $T > K$. We use the shortcut notation $\{\ell\} := \{1, \dots, \ell\}$ for any $\ell \in \mathbb{N}$. Associate to every arm $k \in \{K\}$ a discrete, stationary-time stochastic process $\left(X_t^k\right)_{t\in\mathbb{N}}$ that is defined through its canonical version over the probability space $\left(\Omega_k, \mathcal{B}_k, \mathbb{P}^k\right)$, where, with a slight abuse of notation, we denote $\Omega_k = \mathcal{X}^{\mathbb{N}}$, $\mathcal{B}_k := \mathcal{B}\left(\mathcal{X}^{\mathbb{N}}\right)$ (i.e. Borel sigma algebra over $\mathcal{X}^{\mathbb{N}}$). For the process $\left(X_t^k\right)_{t\in\mathbb{N}}$ we denote also the canonical filtration $\mathcal{F}_i^k := \sigma\{X_j^k, j \leq i\}$. We assume each process to be weakly stationary (i.e., such that $\mathbb{E}\left[X_t^k\right]$ does not change with time t). To give a proper probabilistic model for the stochastic bandit problem, we consider the probability space $(\Omega, \mathcal{A}, \mathbb{P})$, where $\Omega = \Omega_1 \times \ldots \times \Omega_K$; $\mathcal{A}$ is the $K-$product of Borel $\sigma-$algebras $\mathcal{B}_k$ and $\mathbb{P}$ is a joint probability measure for which $k-$th marginal is the measure $\mathbb{P}^k$ on $\mathcal{B}_k$, as defined above. We refer to measure $\mathbb{P}$ as a stochastic bandit (or bandit instance). Note that we do not assume that the measure $\mathbb{P}$ is a product measure on $\mathcal{A}$; in general this model includes possible dependencies between outputs from different arms. We set $\mathcal{F}_i := \sigma\{X_j^{\{K\}} : j \leq i, k \in \{K\}\}$ as the canonical filtration generated by *all* stochastic processes $(X_t^a)_{t\in\mathbb{N}}, a \in \{K\}$. Let $(I_t)_{t\in\mathbb{N}} : \Omega \mapsto \{K\}$ be any map. Define $\tilde{\mathcal{F}}_t^{I\cdot}$ as the filtration that tracks the series of selected arms $(I_t)_{t\geq 1}$ and corresponding outputs of the process $X_t^{I_t}$, formally $\tilde{\mathcal{F}}_t^I = \sigma\Big(I_1, X_1^{I_1}, I_2, X_2^{I_2}, \dots, I_{t-1}, X_{t-1}^{I_{t-1}}\Big)$. Notice that $\tilde{\mathcal{F}}_0^I = \{0, \Omega\}$. A map $I_t : \Omega \mapsto \{K\}$ is called a *strategy* if for any $t \geq 1$, I_t is $\mathcal{F}_{t-1}^{I_t}$ measurable. This technical assumption ensures that the learner's strategy depends only on the decisions from the observed history. Notice that from Theorem 3, p.174 in Shirayev (1996), it follows that I_t can be represented as a measurable function such that $I_t : \{I_s, X_s^{I_s}\}_{s\leq t-1} \mapsto \{K\}$. For a bandit instance $\mathbb{P}$ and strategy $(I_t)_{t\geq 1}$, let $\mathbb{P}_{\mathcal{F}_t^I} := \mathbb{P}\left[\cdot|\mathcal{F}_t^I\right]$ be a regular conditional probability of $\mathbb{P}$ given $\mathcal{F}_t^I$. We assume all the expectations and distributional characteristics to be taken under the bandit instance $\mathbb{P}$. Therefore, we denote $\mu_k = \mathbb{E}\left[X_t^k\right]$ and also $\mu_* := \max_{k\in\{K\}} \mu_k$ for the arm with the biggest average reward. We refer to $* = \underset{k\in\{K\}}{\operatorname{Arg\,Max}}\, \mu_k$ as the arm with the *highest stationary mean* or the best arm. Define $A' := \{k \in \{K\}, \mu_k < \mu_*\}$ as the set of suboptimal arms and write $\Delta_k := \mu_* - \mu_k$ for the average regret of playing suboptimal arm $k \in A'$ in one round.

For each time step $t \leq T$, based on their strategy $(I_t)_{t\geq 1}$, the learner chooses the arm $I_t \in \{K\}$ and receives an output of the stochastic process $X_t^{I_t} \sim \mathbb{P}_t^{I_t}$. As a performance measure of the learner's strategy I on the bandit instance $\mathbb{P}$ over T rounds, we recall (see 1) that we consider the notion of the pseudo-regret as

$$R_{\mathbb{P}}(I,T) = \mathbb{E}\left[\sum_{t=1}^{T} \mu_* - \mu_{I_t}\right] = T\mu_* - \sum_{t=1}^{T} \mathbb{E}[\mu_{I_t}] = \sum_{k=1}^{K} \Delta_k \mathbb{E}[N_k(T)], \tag{4.1}$$

where, for every $k \in \{K\}$, $N_k(T) = \sum_{t=1}^{T} \mathbf{1}\{I_t = k\}$ denotes the number of times the learner chooses the arm k over T rounds, and the expectation is taken under bandit instance $\mathbb{P}$. Next, we give the motivation of the choice of this type of regret in the analysis in particular to the dependent setting.

4.2.1 Different notions of regret

In the classical i.i.d. case, the pseudo-regret of the strategy $I = (I_t)_{t\geq 1}$, $R_{\mathbb{P}}(I,T)$ as defined in Equation (4.1) coincides with the more standard notion of *expected regret*, which is defined with respect to

noise both in the strategy and in the sample. More precisely, the expected regret is defined as

$$\overline{R}_{\mathbb{P}}(I,T) := \max_{a\in\{K\}} \mathbb{E}\left[\sum_{t=1}^{T}\left(X_t^a - X_t^{I_t}\right)\right] = \sum_{t=1}^{T}(\mu_* - \mathbb{E}\big[X_t^{I_t}\big]). \tag{4.2}$$

In the i.i.d. case $T\mu_*$ is an upper bound on the expected reward of *any* strategy. This is attained for the "oracle" strategy which always pulls the best arm. *Neither* of these two facts holds in the setting when the distribution of the arms are non i.i.d. Namely, because *both* the strategy choice $\{I_t = k\}$ at time t and the outcome of the arm k, X_t^k depend on the past, we have $\mathbb{E}[X_t^{I_t}] = \sum_{k=1}^{K}\mathbb{E}\left[\mathbb{I}_{I_t=a}X_t^{I_t}\right] = \sum_{k=1}^{K}\mathbb{E}\left[\mathbb{I}_{I_t=k}X_t^{k}\right] \neq \mathbb{E}[\mu_{I_t}]$, and therefore

$$\overline{R}_{\mathbb{P}}(I,T) = T\mu_* - \sum_{t=1}^{T}\mathbb{E}\left[X_t^{I_t}\right] \neq R_{\mathbb{P}}(I,T) = T\mu_* - \sum_{t=1}^{T}\mathbb{E}[\mu_{I_t}].$$

Furthermore, there exist (oracle) "arm switching" strategies exploiting dependencies that have significantly higher expected rewards[1] than $T\mu_*$ over the course of T rounds (see Example 1 and Example 3 in Ortner et al. (2014)).

Bandit instances with non-i.i.d. observations were the object of recent attention, mostly under a (frozen) Markovian assumption about the underlying stochastic process. In this context, the term "frozen" means that the distribution $\mathbb{P}^a$ of the underlying arm a does not change when the underlying arm has not been played after round t. A Markovian setting for the pseudo-regret analysis was considered in Anantharam et al. (1987); a similar setting was recently studied in Ortner et al. (2014). A type of UCB-strategy for the frozen Markovian rewards was considered in Tekin and Liu (2010). Pseudo-regret upper bounds for the processes that exhibit a type of weak-dependency condition (namely φ–mixing processes) were analyzed by Audiffren and Ralaivola (2015). However, in the latter work, the relation of pseudo-regret to the *expected* regret was not addressed. The last point, as well as the notion of the regret with respect to a (switching arm) policy, was introduced and discussed by Grünewälder and Khaleghi (2017). In this work the authors argued that the expected regret with respect to a policy is the more natural notion if the observed outcomes are direct rewards in the non–i.i.d. scenario. Grünewälder and Khaleghi (2017) provided a UCB type of algorithm for a highest stationary mean identification in a jointly $\phi-$mixing bandit instance. They analysed the pseudo-regret (4.1) while pointing out the approximation of the expected regret by means of the pseudo regret. Furthermore, the approximation bounds are covering both issues (Propositions 3 resp. 11 of Grünewälder and Khaleghi, 2017): they bound the difference between μ_* and the expected reward of the best strategy, with regard to the difference $\left|R_{\mathbb{P}}(I,T) - \overline{R}_{\mathbb{P}}(I,T)\right|$. However, the first bound is linear in T and the second is linear in K, so that the approximation bounds can be of a larger order than the bound on the pseudo-regret itself when K and/or T grows.

The question of interest is to analyze the pseudo-regret in the setting that is described by a weak-dependency assumption of a general kind. In this setting we want to find the policy that targets the process with the highest stationary mean with the pseudo-regret as performance measure. Notice that this problem is easier then finding the optimal policy among all switching arm policies (see Example 1 in Ortner et al. (2014), which shows that the optimal switching arm policy is in general different from the optimal highest average mean policy and has smaller regret). In such a scenario the proper performance measure would be the expected regret $R_{\mathbb{P}}(I,T)$. However, pseudo-regret can be considered a proper proxy for the regret, and, in some scenarios of fast–decaying correlations (see Grünewälder and Khaleghi (2017)), the highest mean policy would be a good approximation of the switching arm counterpart.

An example where the use of the pseudo-regret is of interest and is a good proxy for the regret is the setting of *delayed* rewards. First, a minor variation of the setting which is considered would be to

[1] This second issue vanishes if fixed-arm strategies are the only admissible competitors for the regret.

consider a stochastic process $(Y_t)_{t\geq 1}$ with $Y_t^{I_t} = X_{t+\tau}^{I_t}$ and τ as a fixed delay. In this scenario, the delayed reward is still different from the observation but comes from the same stream and is therefore not independent. However for a very large delay τ we get that $\mathbb{E}\big[Y_t^{I_t}|\mathcal{F}_{t-1}\big] = \mathbb{E}\Big[X_{t+\tau}^{I_t}|\mathcal{F}_{t-1}\Big] \approx \mu_{I_t}$ for a weakly-dependent process with dependencies that vanish over time (see section 4.2.2 for a precise definition) and admissible policy $(I_t)_{t\geq 1}$. In this case, because of fading correlation assumptions, a sample average $\frac{1}{\ell}\sum_{k=1}^{\ell} X_{t+k\tau}$ is a good proxy for the μ_{I_t} when τ is big. In a different setting the reward is actually $X_t^{I_t}$ but is only observed after a delay τ. In other words the decision I_t is $\mathcal{F}_{t-\tau}$ measurable.

4.2.2 Weak dependency (mixing) assumption

We use the definition of the functional weak-dependency assumption (as in Chapter 2) with respect to the collection of $\{K\}$ stochastic processes. Here we work under same $\mathcal{C}-$weak-dependency assumption as that given by Definition 2.2.2 in Chapter 2. Note that this notion is similar to a type of mixingale condition (see Dedecker and Merlevede (2015),Mc Leish (1975)) that includes, in particular, the settings of ϕ–mixing (Kontorovich and Ramanan (2008)), uniform τ–mixing (Wintenberger (2010)) and φ mixing (Maume-Deschamps (2006)). Below, we give the formal definition of the $\mathcal{C}$–mixing bandit instance.

Definition 4.2.1. Consider a $K-$armed stochastic bandit instance $\mathcal{B}$ as a distribution $\mathbb{P}$ of a collection of K stochastic processes $(X_t^a)_{a\in\{K\}}$, $t\in\mathbb{N}$ and $a\in\{K\}$ over measurable space

$$(\Omega,\mathcal{F}) = \bigg(\Big(\mathcal{X}^{\mathbb{N}}\Big)^K, \mathcal{A}\bigg), \tag{4.3}$$

where $\mathcal{A} := \mathcal{B}\Big(\big(\mathcal{X}^{\mathbb{N}}\big)^{\otimes K}\Big)$, being the $\sigma-$field generated by the cylinder sets over $\big(\mathcal{X}^{\mathbb{N}}\big)^{\otimes K}$ and measure $\mathbb{P}$ such that it has marginals $\mathbb{P}^{(a)}$ over $\big(\mathcal{X}^{\mathbb{N}}, \mathcal{B}\big(\mathcal{X}^{\mathbb{N}}\big)\big)$ that are the distribution of the stochastic process $(X_t^a)_{a\in\{K\},t\in\mathbb{N}}$. We say that the $K-$ armed bandit problem is $\mathcal{C}-$*weakly mixing* (or simply weakly mixing while posing the dependency on the class $\mathcal{C}$ implicit) if, for every $a\in\{K\}$, the mixing coefficient

$$\phi_{\mathcal{C}}(k) := \sup_{i\in\mathbb{N},\varphi\in\mathcal{C}_1} \big\|\mathbb{E}\big[\varphi\big(X_{i+k}^a\big)|\mathcal{F}_i^a\big] - \mathbb{E}\big[\varphi\big(X_{i+k}^a\big)\big]\big\|_{L_\infty(\mathbb{P})} \tag{4.4}$$

is such that $\lim_{k\to\infty}\phi_{\mathcal{C}}(k) = 0$ where $\mathcal{F}_i^a := \sigma(X_u^a : u\leq i)\subset\mathcal{A}$.

Assumption 1. We assume additionally that the identity function $\mathbb{I}:\mathcal{X}\mapsto\mathbb{R}$ belongs to the class $\mathcal{C}_1$ and use this in the following to derive the upper bounds for the conditional expectation component. For a bandit instance $\mathbb{P}$, this implies that we have $\big\|\mathbb{E}\big[X_{i+k}^a|\mathcal{F}_i^a\big] - \mathbb{E}\big[X_{i+k}^a\big]\big\|_{L^\infty(\mathbb{P})} \leq \phi_{\mathcal{C}}(k)$ for every $a\in\{K\}$. We use the latter condition in the probabilistic toolbox for the control of conditional expectations in the proof.

Remark 4.2.2. It is easy to see that if we consider the past-sigma algebra generated by the observations of the algorithm $(I_t)_{t\geq 1}$ (i.e. $\mathcal{F}_i^I = \sigma\{I_1, X_1^{I_1}, I_2, X_2^{I_2}, \ldots, I_{s-1}, X_{i-1}^{I_{i-1}}\}\subset\mathcal{A}$), then for every i, by the tower property and Jensen's inequality we have:

$$\begin{aligned}\big\|\mathbb{E}\big[X_{i+k}^a - \mathbb{E}\big[X_{i+k}^a\big]|\mathcal{F}_i^I\big]\big\|_\infty &\leq \big\|\mathbb{E}\big[\mathbb{E}\big[X_{i+k}^a - \mathbb{E}\big[X_{i+k}^a\big]|\mathcal{F}_i^a\big]|\mathcal{F}_i^I\big]\big\|_\infty \\ &\leq \big\|\mathbb{E}\big[X_{i+k}^a - \mathbb{E}\big[X_{i+k}^a\big]|\mathcal{F}_i^a\big]\big\|_\infty \leq \phi_{\mathcal{C}}(k).\end{aligned}$$

Examples of $\mathcal{C}$ weak-mixing processes

Firstly we consider the examples of bandit instances with $\mathbb{P}$ being a product measure. That is, the correspondent real-valued processes $(X_t^a)_{t\in\mathbb{N},a\in\mathcal{A}}$ are stochastically independent for $a\in\{K\}, t\in\{T\}$.

The independent noise process naturally satisfies condition 2.2.2 because $\phi_{\mathcal{C}}(k) = 0$ for all $k \geq 1$. The auto-regressive process of order 1, $X_i = \rho X_{i-1} + \xi_i$, where ξ_i is some bounded i.i.d. noise process is geometrically weak-mixing with rate $\phi_k = \exp\left(-k \log\left(\rho^{-1}\right)\right)$, provided $\rho < 1$. A moving-average process of a finite order $q \in \mathbb{N}$, of the form $W_i = \mu + \sum_{j=0}^{q} \theta_j \psi_{i-j}$, for $i \in \mathbb{Z}$, where $(\psi_i)_{i\in\mathbb{Z}}$ is a sequence of bounded i.i.d. random variables and $(\theta_j)_{0\leq j\leq q}$ is also geometrically weak-mixing (see Dedecker and Merlevede (2015) for general example of process in Banach spaces), provided a certain assumption on the sequence $(\theta_j)_{0\leq j\leq q}$ holds (see for example in Canda (1974), also in Rosenblatt (2000) for a big overview on the mixing properties of linear processes). Furthermore, every recurrent aperiodic finite-state Markov chain can be proven as geometrically weak-mixing with rate $\phi(k) \leq \exp\left(-k \log \lambda^{-1}\right)$, where λ is the second-largest eigenvalue of the transition matrix of the Markov chain. Examples of polynomially weak-mixing processes include several types of Metropolis-Hastings independent samplers in which the proposal distribution does not have a lower bounded density; for such an example we refer to Jarner and Roberts (2002).

4.2.3 Concentration toolbox

Our main technical toolbox is a general type of high probability maximal Hoeffding-type concentration inequality that controls the deviations of the random sum of a real stationary stochastic process $(X_t)_{t\in\mathbb{N}}$. The result is due to Peligrad et al. (2007), and we provide it below for completeness.

Theorem 4.2.3 (Proposition 2 in Peligrad et al. (2007))**.** *Let $(Y_t)_{t\in\mathbb{N}}$ be a stationary real-valued centered process; define $S_n := \sum_{i=1}^{n} Y_i$ and $S_n^* = \max_{i\leq n}|S_n|$. For $t \geq 0$, we have that the following inequality holds:*

$$\mathbb{P}(S_n^* \geq t) \leq 4\sqrt{e}\exp\left(-t^2/2n(\|Y_1\|_\infty + 80\delta_n)^2\right),$$

where $\delta_n = \sum_{j=1}^{n} j^{-\frac{3}{2}} \|\mathbb{E}[S_j|\mathcal{F}_0]\|_\infty$ with $\mathcal{F}_0 = \sigma(Y_0)$. In terms of deviation bounds this is equivalent to say that for any $\delta > 0$ we have that with probability at least $1-\delta$, it holds that

$$S_n^* \leq \sqrt{n}(\|Y_1\|_\infty + 80\delta_n)\sqrt{2\log\left(\frac{\mathcal{A}}{\delta}\right)},$$

where $\mathcal{A} = 4\sqrt{e}$.

For a stationary weak-mixing process $(X_t)_{t\in\mathbb{N}}$ that satisfies assumption 2.2.2, we can apply Theorem 4.2.3 by setting $Y_t := X_t - \mathbb{E}[X_t]$, using the obvious fact that $|S_n| \leq S_n^*$ and $\|\mathbb{E}[S_j|\mathcal{F}_0]\|_\infty \leq \sum_{k=1}^{j}\|\mathbb{E}[X_j|\mathcal{F}_0]\|_\infty \leq \sum_{k=1}^{j}\phi_{\mathcal{C}}(k)$. From the definition (2.2.2) under **Assumption 1** we deduce the following proposition.

Proposition 4.2.4. *For a stationary real-valued weak-mixing process $(X_t)_{t\in\mathbb{N}}$ with rate $\phi_{\mathcal{C}}(\cdot)$, S_n being partial sum as above and $\mu = \mathbb{E}[X_t]$, for any $\delta \in [0,1)$ with probability at least $1-\delta$ it holds:*

$$\left|n^{-1}S_n - \mu\right| \leq \left(1 + 80\sum_{j=1}^{n} j^{-\frac{3}{2}} \sum_{k=1}^{j}\phi_{\mathcal{C}}(k)\right)\sqrt{\frac{2\log\left(\frac{\mathcal{A}}{\delta}\right)}{n}}. \tag{4.5}$$

Remark 4.2.5. Notice that the statement of Proposition 4.2.4 trivially extends to the case when instead of $(X_t)_{t\in\mathbb{N}}$, we consider $(X_{kt})_{t\in\mathbb{N}}$ for $k \in \mathbb{N}$ (i.e. a sequence of random variables with gaps of fixed

size). Namely, we have that with probability at least $1-\delta$:

$$\left| n^{-1} \sum_{t=1}^{n} X_{kt} - \mu \right| \leq \left(1 + 80 \sum_{j=1}^{n} j^{-\frac{3}{2}} \sum_{\ell=1}^{j} \phi_{\mathcal{C}}(k\ell)\right) \sqrt{\frac{2\log\left(\frac{A}{\delta}\right)}{n}}. \tag{4.6}$$

The latter result enables the deviation control of the estimate of the mean in the scenario when samples are taken at given sequence of timepoints.

Remark 4.2.6. Theorem 4.2.4 implies that there is a contamination term in the typical Hoeffding's concentration bound. If $\phi_{\mathcal{C}}(t) \leq t^{-\alpha}$ with $\alpha > 1/2$, then approximating the sum by the integral we get

$$\sum_{j=1}^{n} j^{-\frac{3}{2}} \sum_{k=1}^{j} \phi_{\mathcal{C}}(k) \leq c_\alpha \sum_{j=1}^{n} j^{-\frac{1}{2}-\alpha},$$

where c_α is some constant that depends only on α. The last partial sum is convergent for all $\alpha > \frac{1}{2}$. Thus, from Proposition 4.2.4, we have that, for any $\delta > 0$ with probability at least $1-\delta$ w.r.t. distribution $\mathbb{P}$ of stochastic process $(Y_t)_{t\in\mathbb{N}}$, it holds:

$$\left| n^{-1} S_n - \mu \right| \leq (1+M) \sqrt{\frac{2\log\left(\frac{A}{\delta}\right)}{n}}, \tag{4.7}$$

where $M = 80 c_\alpha \sum_{j=1}^{+\infty} j^{-\frac{1}{2}-\alpha} < \infty$ and c_α is a constant that depends only on α. We refer to this type of weak dependence as to a *fast mixing scenario*. If $0 < \alpha < \frac{1}{2}$, then $\sum_{j=1}^{n} j^{-\frac{3}{2}} \sum_{k=1}^{j} \phi_{\mathcal{C}}(k) \leq C_\alpha n^{\frac{1}{2}-\alpha}$, which implies that for partial sums S_n for any $\delta > 0$, it holds with probability at least $1-\delta$ that

$$\left| n^{-1} S_n - \mu \right| \leq \sqrt{\frac{2\log\frac{A}{\delta}}{n} C_\alpha} + \sqrt{\frac{2\log\frac{A}{\delta}}{n^{2\alpha}}}. \tag{4.8}$$

4.3 Main Algorithm ($\mathcal{C}-$Mix UCB) and main regret upper bounds

The learning algorithm we present is conceptually based on the celebrated IMPROVED-UCB algorithm by Auer and Ortner (2010) (see also Evan-Dar et al. (2006) and Perchet and Rigollet (2013) for other schemes of sequential algorithms which are based on the action-elimination policies). We also distinguish an essential difference between the cases in which all arms are polynomially mixing with exponent smaller than $1/2$ *(slow $\mathcal{C}$-mixing)* and the case which all arms are mixing sufficiently fast, typically polynomially mixing with exponent larger than $1/2$, or exponentially mixing, such that the mixing coefficients for each arm are either summable or such that partial sums of order n diverge at speed not faster than $\mathcal{O}(\sqrt{n})$ *(fast $\mathcal{C}$-mixing)*.

The $\mathcal{C}-$MIX UCB algorithm works as follows. Given the number of rounds T, the algorithm divides it into the epochs of nearly exponentially increasing number of pulls t_s for every epoch s. At the given epoch s, the samples from each of the active arms are collected during the sequence of times with a constant gap b_s. This gap b_s is equal to the number of active arms in the epoch s. At the end of the epoch, the estimators of the mean and confidence width of each arm are computed, and the arms for which the lower confidence bound is small (in comparison with the upper confidence bound of the best arm in the given epoch s) are "eliminated". That is, no pulls from these arms are considered in the next epoch. The algorithm then proceeds to the next epoch, and the actions are repeated. For the regret analysis of $\mathcal{C}-$MiX-UCB, we assume the dependence on the sequence of decision rules which is output by the algorithm implicit and write $R_{\mathbb{P}}(T)$ skipping the dependence on the strategy $(I_t)_{t\geq 1}$.

Note that the start τ_s of epoch $s \geq 1$ is random and that an event $\{\tau_s \leq t\}$ belongs to the $\sigma-$field $\mathcal{F}^{I.}_{t-1}$. The sampling scheme of the epoch itself depends on the number of arms not eliminated during previous epochs (and is therefore random); however, given this information, (mathematically represented as the σ-algebra $\mathcal{F}^I_{\tau_s}$), the sampling scheme is deterministic. In such a learning scheme, to be able to use concentration inequalities to estimate the mean of each arm from the samples collected during the current epoch s, one must ensure that the process $\tilde{X}^a_t := \left(X^a_{\tau_s+t}\right)_{t\in\{T\}}$ is $\mathcal{C}$–mixing *conditionally to* $\mathcal{F}_{\tau_s}$, whenever the process X^a_t is $\mathcal{C}$–mixing.

Note that this problem does not arise in the basic i.i.d. bandit instance, as all characteristic properties of the independent bandit instance $\mathbb{P}$ carry over to the distribution of the process $\tilde{X}^a_t$ conditioned on the $\mathcal{F}^I_{\tau_s}$. The following statement justifies that the $\mathcal{C}-$mixing property transforms from any bandit instance $\mathbb{P}$ to its conditional measure $\mathbb{P}_{\mathcal{F}^I_{\tau_s}}[\cdot]$ with respect to the strategy of $\mathcal{C}-$MIX UCB algorithm.

Proposition 4.3.1. *Consider a $K-$armed stochastic bandit instance $\mathbb{P}$ over the space $(\Omega, \mathcal{F})$ with $\Omega = \Omega_1 \times \ldots \times \Omega_K$ and $\mathcal{F} = \mathcal{B}\left(\mathcal{X}^{\mathbb{N}}\right) \times \ldots \times \mathcal{B}\left(\mathcal{X}^{\mathbb{N}}\right)$ which is $\mathcal{C}-$mixing with rate $\phi_{\mathcal{C}}(\cdot)$ and that **Assumption 1** holds. For any $s \in \mathbb{N}$, denote τ_s as the start of epoch s as given in Algorithm 2. Denote $\mathcal{F}^I_{\tau_s}$ to be the $\sigma-$algebra generated by the stopping time τ_s and $\mathbb{P}[.|\mathcal{F}_{\tau_s}]$ to be the regular conditional distribution of the process $X^{\{K\}}_t$ conditional to $\mathcal{F}_{\tau_s}$. Then, for every $a \in \{K\}$, it holds $\mathbb{P}-$a.s. that the process $\tilde{X}^a_t = X^a_{\tau_s+t}$ is $\mathcal{C}$-mixing with rate bounded by $2\phi_{\mathcal{C}}(t)$ under $\mathbb{P}[.|\mathcal{F}_{\tau_s}]$.*

Remark 4.3.2. It is easy to check that if the marginal processes $(X^a_t)_{t\leq T}$ are stochastically independent over $a \in \{K\}$, then it is sufficient to prove the above property for arbitrary (one) process $(X^a_t)_{t\leq T}$, $a \in \{K\}$.

In the following, we assume an a priori upper bound on the mixing rate of the mixing bandit instance $\mathbb{P}$ to be $\phi_{\mathcal{C}}(t)$; this rate (more precisely the exponent α) will determine the epoch's size of the $\mathcal{C}-$MIX UCB as well as the pseudo-regret upper bounds in the case of slow mixing rates.

Remark 4.3.3. We remark that for our analysis, it is sufficient to choose the upper bound on the last epoch $s_{\text{end}} := \lfloor \frac{1}{2}\log\left(\frac{AT}{32}\right)\rfloor$. Indeed, in Algorithm 2 one can check by direct computation that for all s, $T_{s,1} > T_{s,2}$ and furthermore $\log\left(\mathcal{A}T\theta^2_s\right) > 1$ for all $s \leq s_{\text{end}}$. Taking this into account, by plugging in s_{end} into $T_{s,1}$ we obtain that at this epoch $T_{s_{\text{end}}} > T$.

4.3.1 Fast mixing scenario

In the *fast mixing* scenario, for which, as mentioned before, $\phi_{\mathcal{C}}(t) = t^{-\alpha}$ with $\alpha > \frac{1}{2}$ we make use of concentration inequality (4.7) from Remark 4.2.6 in the IMPROVED-UCB learning scheme. This algorithm was originally presented in Auer and Ortner (2010) and rigorously analyzed by Perchet and Rigollet (2013). The latter considers sequential arm elimination during epochs of increasing lengths, in which the arm's pulling sequences in each epoch is an arbitrary deterministic sequence. For example, we can pull arms in the circular way as represented in the following section of the slow mixing regime. The latter means that the time-gaps between the two consecutive pulls of one arm equals the number of arms active in epoch s.

For any epoch s, let τ_s be its random start. By Proposition 4.3.1, for any arm $a \in \{K\}$, the samples from arm a that are collected during the epoch s satisfy (conditionally with respect to the random start of the epoch s) the $\mathcal{C}-$mixing property with rate $2\phi_{\mathcal{C}}(t)$. Therefore, conditioned to the random start of the epoch s, we can directly use the concentration inequality (4.7), replacing its counterpart for the case of product measure $\mathbb{P}$ as in the proof of Theorem 3.1 in Auer and Ortner (2010) for each given epoch s.

We observe that, apart from the multiplicative constant in the concentration inequality, the contamination term has no other influence on the concentration rate. Therefore, the analysis of Auer and Ortner (2010) can be repeated directly for the *fast* $\mathcal{C}$–mixing processes. This directly implies the following problem-dependent regret upper bound.

Parameters: K and T

Initialization: $\mathcal{A} = 4e^{1/2}$, $c_0 = ((1-\alpha)(1/2-\alpha))^{-1}$, $c_1 = \left(\frac{(1-\alpha)(1/2-\alpha)}{80}\right)^{\frac{2}{1-2\alpha}}$, $c_3 = 12800c_0$, $s = 1$, $\tau_0 = 1$, $B_0 = \{K\}$

while $t \leq T$ **do**

$\theta_s = 2^{-s}$, $t_s := \left(32c_1^{-1}\theta_s^{-2}\log\left(\mathcal{A}T\theta_s^2\right)\right)^{\frac{1-2\alpha}{2\alpha}}$, $b_s = |B_s|$;

Select number of pulls T_s

$$T_s = T_{s,1} := \left\lceil \frac{32\log\left(\mathcal{A}T\theta_s^2\right)}{\theta_s^2} \right\rceil \text{ for } b_s \geq t_s;$$

$$T_s = T_{s,2} := \left\lceil \frac{1}{b_s}\left(\frac{c_3\log\left(\mathcal{A}T\theta_s^2\right)}{\theta_s^2}\right)^{\frac{1}{2\alpha}} \right\rceil, \text{ for } b_s < t_s.$$

for $\ell \in \{0, \ldots, T_s - 1\}$, $i \in B_s$ **do**

if $\tau_s + i + \ell b_s > T$ **then**

BREAK

end

else

choose arm i in B_s at a time point $\tau_s + i + \ell b_s$

end

end

Compute

$$\Omega(\theta_s, b_s) = \left(1 + 80\sum_{j=1}^{T_s} j^{-\frac{3}{2}} \sum_{\ell=1}^{j} \phi_{\mathcal{C}}(b_s\ell)\right)\sqrt{\frac{2\log(\mathcal{A}T\theta_s^2)}{T_s}},$$

$$\widehat{\mu}_{i,s} = T_s^{-1}\sum_{t=0}^{T_s-1} X^i_{\tau_s+i+tb_s}$$

$B_s = B_s \setminus \{i \in B_s : \widehat{\mu}_{i,s} + \Omega(\theta_s, b_s) \leq \max_{j\in B_s}\widehat{\mu}_{j,s} - \Omega(\theta_s, b_s)\}$

$s = s + 1$

$\tau_s = \tau_{s-1} + b_s T_s$

end

Algorithm 2: Algorithm C-Mix UCB

Theorem 4.3.4. *The pseudo-regret of the* $\mathcal{C}-$MIX UCB *algorithm for the stochastic bandit instance* $\mathbb{P} = B_{\phi_{\mathcal{C}}(\cdot),K}$, $\phi_{\mathcal{C}}(\cdot)$ *in a fast* $\mathcal{C}-$*mixing bandit scenario is bounded by*

$$R_{\mathbb{P}}(T) \leq (1+M)\sum_{k\in A_\lambda}\Delta_k + \frac{96}{\Delta_k} + \frac{32\log\left(T\Delta_k^2\right)}{\Delta_k} + 64\sum_{k\in A_0\setminus A_\lambda}\frac{1}{\lambda} + T\max_{k\in A_0\setminus A_\lambda}\Delta_k,$$

where $M = 80c_\alpha\sum_{j=1}^{+\infty} j^{-\frac{1}{2}-\alpha}$, $\lambda \geq \frac{e^{\frac{1}{4}}}{2\sqrt{T}}$ *is chosen arbitrarily and* $A_\lambda = \{k \in \{K\} s.t. \Delta_k > \lambda\}$.

Remark 4.3.5. Notice that for any $\lambda \geq \frac{e^{1/4}}{2\sqrt{T}}$, we have that the following holds. In the first sum, the term $\frac{32\log(T\Delta_k^2)}{\Delta_k}$ dominates the bound. Furthermore, if the set $A_0 \setminus A_\lambda$ is non-empty, then term $\frac{64}{\lambda}$ majorates $T\max_{k\in A_0\setminus A_\lambda}\Delta_k$ in the bound. Thus, choosing λ to balance between $\sum_{k\in A_\lambda}\frac{32\log(T\Delta_k^2)}{\Delta_k}$ and $\sum_{k\in A_0\setminus A_\lambda}\frac{64}{\lambda}$, we minimize the bound on $R_{\mathbb{P}}(T)$.

Considering the $K-$armed bandit instance with $\Delta_k = \sqrt{\frac{K\log(K)}{T}}$ for $k = 2, \dots, K$ and optimizing over λ, we obtain the problem-independent regret upper bound analogous to Remark 3.3 by Auer and Ortner (2010).

Theorem 4.3.6. *In the fast mixing scenario, the* $\mathcal{C}$-MIX UCB *policy satisfies the following (problem independent) pseudo-regret upper bound*

$$R_{\mathbb{P}}(T) \leq 125\sqrt{(1+M)KT}\frac{\log(K\log(K))}{\sqrt{\log(K)}}.$$

4.3.2 Slow mixing scenario

In this part we consider the slow mixing scenario. In this case, the mixing rate of the stochastic processes X_t^a for $a \in \{K\}$ is assumed $\phi_{\mathcal{C}}(t) = t^{-\alpha}$ with $\alpha \in (0, 1/2]$. Theorem 4.3.8 provides the problem-dependent upper bound for the pseudo-regret of the $\mathcal{C}-$MIX UCB learning strategy (Algorithm 2) in the case of the slow mixing scenario. In every epoch s algorithm pulls all remaining active arms are pulled cyclically, so that the time gap between two consequent pulls of one arm is equal to the number of active arms in the given epoch. This sequence is deterministic, given the samples and strategies choices until τ_s – the random start of epoch s. Formally this means that random variables τ_s is $\mathcal{F}^I_{\tau_s}$ - measurable. To estimate the mean in the slow mixing scenario before the epoch s we consider samples collected *during the time length of the epoch* s (and not during all the time until the epoch s has started). However, as the epoch's length increases geometrically with power larger than 2, which implies that at the time point of the end of the epoch s, more than the half of the samples of each arm are collected exactly during the epoch s (and not earlier), up to a multiplicative constant 2; this provides the same confidence term $\Omega(\theta_s, b_s)$ as if we were considering all the samples from the beginning.

Remark 4.3.7. We treat the case with $\alpha = \frac{1}{2}$ separately. Using Remark 4.2.6 we deduce that $\sum_{t=1}^{T}\frac{1}{t} \lesssim \log(T)$, so in this case we can apply the same scheme as in Theorem 4.3.4, with $M = \log(T)$. In this case we obtain the regret upper bounds as in the i.i.d. scenario with the only contamination factor of order $\log(T)$.

Theorem 4.3.8 (Pseudo-regret upper bound for the *slow mixing* scenario)**.** . *Consider the* $\mathcal{C}-$ *polynomially mixing bandit instance* $\mathbb{P}$ *with the rate* $\Phi_{\mathcal{C}}(t) \leq t^{-\alpha}$, $\alpha \in (0, 1/2)$. *Then the* $\mathcal{C}-$MIX-UCB*, which returns sequences of decision* I_t*, satisfies the following pseudo-regret upper bound:*

$$R_{\mathbb{P}}(T) \leq 2\sum_{k\in A_\lambda}\max\{c_2\Delta_k^{-1}\max\{\log\left(\mathcal{A}T\Delta_k^2\right), 1\}, 1\} + \tilde{c}(\Delta_{*,\lambda})^{1-\frac{1}{\alpha}}\left(c_3\log\left(\mathcal{A}T\Delta_{*,\lambda}^2\right)\right)^{\frac{1}{2\alpha}}$$
$$+\frac{12}{\sqrt{e}}\sum_{k\in A_0\setminus A_\lambda}\frac{1}{\lambda} + T\max_{k:k\in A_0\setminus A_\lambda}\Delta_k,$$

where all numerical constants are defined as $\mathcal{A} = 4\sqrt{e}$, $c_2 = 64c_0$, $c_0 = ((1-\alpha)(1/2-\alpha))^{-1}$, $c_1 = \left(\frac{(1-\alpha)(1/2-\alpha)}{80}\right)^{\frac{2}{1-2\alpha}}$, $c_3 = 12800c_0$, $c_4 = \frac{1}{1.2\sqrt{2.4}^{\frac{1}{\alpha}-2}-1}$, $\tilde{c} = 2^{-\frac{1}{\alpha}+3}c_4c_3^{\frac{1}{2\alpha}}$ *and* $\Delta_{*,\lambda} = \min_{j\in A_\lambda}\Delta_j$*, whereas* $\lambda > 0$ *can be chosen arbitrarily.*

Remark 4.3.9. Note that with $\lambda \geq \sqrt{\frac{e^{1-1/e}}{T}}$ for $k \in A_\lambda$ we have $1 \leq \log\left(\mathcal{A}T\Delta_k^2\right) \leq \log(T)$ and $\log(T) > 1$. Thus we obtain the following Corollary in terms of the threshold λ and the additive dependency term.

Corollary 4.3.10. *For any choice of λ that satisfies Remark 4.3.9, one has the following upper bound:*

$$R_{\mathbb{P}}(T) \leq \mathcal{O}\left(\sum_{k\in A_0} \frac{\log(T)}{\lambda}\right) + \mathcal{O}\left(\Delta_{*,\lambda}^{\frac{\alpha-1}{\alpha}} \log^{\frac{1}{2\alpha}}(T)\right) + T \max_{k:k\in A_0\setminus A_\lambda} \Delta_k, \tag{4.9}$$

where $\Delta_{*,\lambda}$ is defined as in Theorem 4.3.8.

Remark 4.3.11. From the definition of $\Delta_{*,\lambda}$, it follows that $\Delta_{*,\lambda}^{\frac{\alpha-1}{\alpha}} \leq \lambda^{\frac{\alpha-1}{\alpha}}$. This implies the following upper bound for the pseudo-regret in terms of the threshold λ:

$$R_{\mathbb{P}}(T) \leq \frac{K\log(T)}{\lambda} + \lambda^{\frac{\alpha-1}{\alpha}} \log^{\frac{1}{2\alpha}}(T) + \lambda T.$$

Furthermore, by straightforward comparison of the first two summands, for any choice λ from Theorem 4.3.8, if $\lambda \leq \left(\frac{\log(T)}{K^{\frac{\alpha}{1-2\alpha}}}\right)$, the term $\lambda^{\frac{\alpha-1}{\alpha}} \log^{\frac{1}{2\alpha}}(T)$ dominates the other. Otherwise, if $\lambda > \left(\frac{\log(T)}{K^{\frac{\alpha}{1-2\alpha}}}\right)$, we have that $\frac{K\log(T)}{\lambda}$ is of a larger order.

Analyzing the worst-case scenario for the polynomially weak mixing processes, we obtain the following (problem-independent) upper bound.

Theorem 4.3.12 (Problem-independent upper bound)**.** *Assume we have a polynomially weak-mixing instance $\mathbb{P}$ such that Theorem 4.3.8 holds. Then, the $\mathcal{C}$-*MIX UCB *learning algorithm satisfies the following (instance-independent) pseudo-regret bound:*

$$R_{\mathbb{P}}(T) \leq C_3\sqrt{T}\max\{\sqrt{K\log T}, T^{1/2-\alpha}(\log T)^{\frac{1}{2\alpha}}\},$$

where C_3 is some absolute numerical constant.

When $\alpha \to 0$ the regret bound scales almost linearly with the number of rounds T.

4.4 Problem independent lower bounds for regret in dependent bandit scenario

It is natural to investigate whether the regret upper bounds obtained in the previous section are optimal(i.e. to search for lower bounds on the pseudo-regret $R_{\mathbb{P}}(I,T)$ with respect to any admissible strategy). This question can be addressed by classifying bandit instances $\mathbb{P}$ according to their decay rate of the $\phi_{\mathcal{C}}(\cdot)$ mixing coefficient. Firstly, recall that in the fast mixing scenario, upper bounds of Theorems 4.3.4 and 4.3.6 match (up to a multiplicative absolute constant) the corresponding problem-independent regrets bounds for stochastic i.i.d. bandits. From Bubeck and Cesa-Bianchi (2012), Bubeck (2010) it is well known that $\sup\inf R_{\mathbb{P}}(T) \geq c\sqrt{TK}$, where the infimum is taken over all admissible strategies, the supremum is taken over all *stochastic independent* bandit instances, and c is some small numerical constant. The latter lower bounds imply that in the broader stochastic fast mixing bandits scenario (which trivially extends independent stochastic bandits), our regret bounds are optimal up to a $\log(T)$ factor and a multiplicative constant that depends on the ℓ_1 norm of the mixing sequence. In the case of problem-dependent lower bounds, it is known (see Auer (2002)) that the $UCB1$ type of strategy is optimal in the stochastic independent bandit case. More precisely, it is known that for any $\varepsilon > 0$, there is no learning strategy such that it holds $R_{\mathbb{P}}(I,T) \leq \sum_{k:\Delta_k>0} \frac{\log(T)}{(2+\varepsilon)\Delta_k}$, uniformly over independent distributions of arms $\left(X_t^k\right)$, $k \in \{K\}$, $t \in \{T\}$. The latter implies that the $\mathcal{C}-$mix UCB algorithm is optimal in the *fast* mixing scenario.

To fill the existing gap, it is interesting to consider the problem of lower bounds for stochastic bandits when all admissible environments are slow mixing. Audiffren and Ralaivola (2015), Grünewälder and

Khaleghi (2017) also analyse the setting of dependent bandits. However, the question of regret lower bounds is not approached there. Below, we provide the problem-independent lower bound that matches (up to a factor of order $\log^{\frac{1}{2\alpha}}(T)$) the regret upper bound in the case of a slow–mixing scenario.

Consider the the set $B_{\phi_{\mathcal{C}},K}$ of all $K-$armed weak-mixing stochastic bandit instances which satisfy Definition 2.2.2 with the correlation decay rate $\phi_{\mathcal{C}}(k)$.

Theorem 4.4.1 (Problem independent lower bound for weak-mixing stochastic bandits)**.** *For any bandit instance $\mathbb{P} \in B_{\phi_{\mathcal{C}},K}$, rate function $\phi_{\mathcal{C}}(k) = k^{-\alpha}$, and $0 < \alpha \leq \frac{1}{2}$ over all admissible learning strategies, the following lower bound on the pseudo-regret holds:*

$$\inf_{(I_t)_{t\geq 1}} \sup_{\mathbb{P}\in B_{\phi_{\mathcal{C}},K}} R_{\mathbb{P}}(I,T) \geq \frac{\left(\sqrt{2}-1\right)^2}{8} T^{1-\alpha}.$$

4.5 Discussion

4.5.1 Learning scenarios with independence regime

The upper bounds for the pseudo-regret in the fast mixing case of Theorem 4.3.4 and Theorem 4.3.6, match the analogous results in the i.i.d. data scenario, up to a multiplicative absolute constant. Even in the cases where the series $\sum_{t=1}^{\infty} \phi_{\mathcal{C}}(t)$ diverges, the influence of the penalization term due to dependence can be bounded by a constant, assuming polynomial rate with $\frac{1}{2} > \alpha$. Moreover, the independence regime regret upper bound can be recovered even under slow mixing scenario (i.e. with $0 < \alpha \leq 1/2$). Namely, from the statement of Theorem 4.3.12, it follows that in the case $K > T^{1-2\alpha}$, the main contribution to the regret upper bound in Corollary 4.3.10 is given by the first term, which matches the well-known upper bound in the independent data scenario for UCB-type algorithms (see, for example, Bubeck and Cesa-Bianchi (2012); also Auer (2002)). Also, in this scenario, for $\alpha \to 1/2$ we recover the optimal problem-independent bound (up to a square root of a logarithmic term in the number of rounds). In the special case $\alpha = 1/2$ we will have the typical UCB bound, contaminated by a multiplicative term $\log(T)$ (from the influence of the sum of dependent coefficients).

4.5.2 Comparison with the known regret upper bounds

The existing literature on the regret analysis for the problem of stochastic bandits with dependent reward observations is relatively scarce. Audiffren and Ralaivola (2015) consider a type of weak-dependent process called a $\tilde{\phi}_{\mathcal{C}}-$mixing. In the setting of the slow mixing scenario, they obtain an asymptotic regret upper bound of order $\widetilde{\Theta}\big(\Delta_{*,\lambda}^{\frac{\alpha-2}{\alpha}} \log^{\frac{1}{\alpha}}(T)\big)$, where the notation $\widetilde{\Theta}(f) = g$ means that there exists $\gamma, \beta > 0$ so that $|f|\log^{\beta}(|f|) \leq |g|$ and $|g|\log^{\gamma}(|g|) \leq |f|$. Comparing the result of the Proposition 1 with the bound of Audiffren and Ralaivola (2015), we observe an improvement in regret upper bounds in several regards. First, our upper bound is not asymptotic in nature; it also depends explicitly on the gaps of *all* suboptimal arms, while Audiffren and Ralaivola (2015) provide an upper bound in terms of the worst gap only (which corresponds to the penalty term $\Delta_{*,\lambda}$). Second, Corollary 4.3.10 gives an asymptotic bound, which (in the case when the penalty term $\Delta_{*,\lambda}^{\frac{\alpha-1}{\alpha}} \log^{\frac{1}{2\alpha}}(T)$ has the largest impact on the bound) is better by a polynomial factor in terms of the smallest gap and in the power of log-term of the time-horizon. Also, the additive penalty term due to dependency from Theorem 4.3.8 does not scale with the number of arms (which improves over the similar results of Audiffren and Ralaivola, 2015 for the particular case of $\varphi-$mixing processes). Lastly, our analysis is provided for a broader class of weak-dependent processes, which, as in its particular case it includes φ–mixing.

Furthermore, comparing our pseudo-regret upper bounds with the result by Grünewälder and Khaleghi (2017) (Theorem 10 therein), we observe that in the fast–mixing scenario the latter scales in the same

way (with a constant factor that depends on the sum of mixing coefficients) and has the same order of magnitude in Δ_k and T. However, our work contributes to the analysis of the slow–mixing scenario (and for a much general class of processes), which was not covered in Grünewälder and Khaleghi (2017) and provides the matching (up to log terms) upper bound, showing optimality in the slow mixing scenario.

Remark 4.5.1. The pseudo-regret upper bounds from Theorem 4.3.8 cannot be obtained by simply using the standard Improved-UCB algorithm while plugging-in variations of conditional Hoeffding's concentration inequalities with "worse" deviation rates (see, e.g., Bubeck et al. (2013) for heavy-tailed bandits). With such an approach, one gets a penalty term with the worse rate inside the sum over suboptimal arms, so that the pseudo-regret scales linearly with the number of arms. The surprising effect of additive contamination is specific to the weak-dependent scenario and the proposed strategy, exploiting the knowledge of mixing coefficients and the number of arms in each epoch. For comparison, notice that we can still use the fast mixing results in the slow mixing scenario, because T is finite and we can take there $M = \sum_{t=1}^{T} \phi_{\mathcal{C}}(t)$ (now growing with T). Comparing the problem-independent upper bound of Theorem 4.3.6 (standard Improved-UCB) with that of Theorem 4.3.12 (Algorithm 2), we observe that Algorithm 2 gives better bounds. Namely, applying the standard Improved-UCB learning scheme in the slow–mixing regime results in the regret upper bound in Theorem 4.3.6 being impacted by a multiplicative scaling factor $1 + M \sim \sum_{t=1}^{T} \phi_{\mathcal{C}}(t) \sim T^{1/2-\alpha}$. Disregarding the influence of the logarithmic terms, this gives a bound of order $T^{1-\alpha}\sqrt{K}$. This is worse than the bound of Theorem 4.3.12, which does not have the scaling factor in the number of arms in the corresponding term.

4.5.3 Dependent setting with delays

We shortly highlight another setting, where the developed theory of pseudo-regret analysis is of interest. Namely, we consider the case when there is intristic delay between dependent observations and actions. Let $\tau > 0$ be some integer number. Assume that because of various constraints, the learner makes the choice I_t based not on immediate history, but on the outcomes observed up until time $t - \tau$. This is a specific case of the so-called delayed bandit problem, first studied for independent outcome observations by Guha et al. (2010a). The effect of the delay in this setting is an *additive* penalty (depending linearly on τ) for the regret (see Joulani et al., 2013; and Desautels et al. (2014) for the case of Gaussian delayed rewards). If we consider the case of arbitrary data sequences that are fixed in advance (i.e. the adversarial bandit problem), Neu et al. (2013) showed a regret upper bound increased by a multiplicative factor in τ with respect to the standard case (see also Joulani et al. (2013) for the more general problem with random delay times). We present here the intermediate position of the random weakly dependent setting in the delayed feedback bandit problem by considering the cases when the pseudo-regret is a good proxy for the regret. Formally, we define an admissible $\tau-$delayed policy $I = (I_t)_{t\geq 1}$ as a function taking values in $\{K\}$ as follows:

$$I_t = \begin{cases} \text{choose arm randomly} & \text{if } t < \tau; \\ I_t\Big(X_{t-\tau}^{I_{t-\tau}}, \ldots, X_1^{I_1}, I_1, \ldots, I_\tau\Big) & \text{if } t \geq \tau. \end{cases} \tag{4.10}$$

In other words, by putting $Y_i := X_i^{I_i}$ for $i \geq 1$ and recalling that $\mathcal{F}_t^I := \sigma\{Y_1, I_1, \ldots, Y_t, I_t\}$ we assume I_t to be $\mathcal{F}_{t-\tau}^I$ measurable. We now show that in the delayed feedback setting, the pseudo-regret is a good approximation of the expected regret if the delay is large. Consider the expectation of the sample rewards $\mathbb{E}\big[\sum_{t=1}^{T} X_t^{I_t}\big]$. By using the tower property and the definition of weak-$\mathcal{C}$-mixing bandit instance

we have:

$$\mathbb{E}\left[\sum_{t=1}^{T} X_t^{I_t}\right] = \sum_{t=1}^{T}\sum_{k=1}^{K} \mathbb{E}\left[X_t^{I_t}\mathbb{I}_{I_t=k}\right]$$
$$\geq \sum_{t=1}^{T}\sum_{k=1}^{K} \mathbb{E}[(\mu_{I_t} - \phi_{\mathcal{C}}(\tau))\mathbb{I}_{I_t=k}] = \sum_{t=1}^{T} \mathbb{E}[\mu_{I_t}] - \phi_{\mathcal{C}}(\tau)T.$$

By symmetry, we can apply the same reasoning and get the reverse bound. By uniting these contributions, we get the two-sided control over $\mathbb{E}\big[\sum_{t=1}^{T} X_t^{I_t}\big]$:

$$\left|\mathbb{E}\left[\sum_{t=1}^{T} X_t^{I_t}\right] - \sum_{t=1}^{T} \mathbb{E}\left[\mu_{I_t}\right]\right| \leq \phi_{\mathcal{C}}(\tau)T, \tag{4.11}$$

which, together with the definitions of expected regret and pseudo-regret implies that for any strategy π_τ (i.e., a choice of decision functions I_t adapted to the filtration $\mathcal{M}_{t-\tau}$)

$$\left|R_{\tau,\mathbb{P}}(\pi_\tau) - \overline{R}_{\tau,\mathbb{P}}(\pi_\tau)\right| \leq \phi_{\mathcal{C}}(\tau)T, \tag{4.12}$$

where the index τ indicates the τ-delayed setting.

Thus is the delayed setting, if the coefficient $\phi(t)$ is small the pseudo-regret is a good proxy for the expected regret $R_{\tau,\mathbb{P}}(\pi_\tau, T)$. In general, however, the question of true regret measure in the setting with non-i.i.d. observations remains to be open.

4.6 Conclusions

In this chapter, we have examined the extension of the stochastic bandit problem to the case where the bandit instance satisfies a weak-dependency assumption of a general kind. It characterizes the decay of correlations between the past and the future of the process and is measured by $\phi_{\mathcal{C}}-$mixing coefficients. Using the $\mathcal{C}$-MIX UCB Algorithm, in many scenarios we recover (i.e., in the "fast mixing" case where the mixing coefficients have either exponential or polynomial decay with power $\alpha > \frac{1}{2}$) we (up to an absolute multiplicative constant which depends on the ℓ_1 norm of the sequence of mixing coefficients), the same pseudo-regret upper bounds as in the independent data scenario.

Furthermore, even in the case where the processes are slowly mixing (i.e., when $\phi_{\mathcal{C}}(t) \sim t^{-\alpha}$ with $\alpha < \frac{1}{2}$), the presented $\mathcal{C}$-Mix UCB algorithm has the regret upper bound incurring only an additive penalty compared with the independent outcomes scenario for problem-dependent upper bound. Under certain conditions on the relation between the number of arms, the time horizon, and the mixing rate, a proper choice of the threshold in the penalty allows recovery of the same regret upper bounds as for independent data observations. In other regimes, our algorithm highlights the surprising effect that the worst-case upper bound does not scale with the number of arms.

An interesting and non-trivial question for further study would be to approach directly the notion of the expected regret in the dependent setting. Despite different attempts (see Grünewälder and Khaleghi (2017) for the approach of approximation of the regret by means of the pseudo regret, Ortner et al. (2014) for the analysis of different notions of the regret in the setting of Markovian outputs) to tackle this problem, the optimal notion of the regret in the setting of dependent arm's observations remains to be unclear.

4.7 Proofs of the main results of Chapter 4

4.7.1 Necessary toolbox for the proof of the main probabilistic result (Theorem 4.3.1)

To prove Proposition 4.3.1, we need an auxiliary probabilistic statement that may be of independent interest. As in Section 4.2, we consider a stochastic bandit instance $\mathbb{P}$ such that for every $a \in \{K\}$, we have $(X_t^a)_{t\in\mathbb{N}}$ is weak-mixing with rate $\phi_{\mathcal{C}}(\cdot)$ and that **Assumption 1** holds. If $(X_t^a)_{t\in\mathbb{N}}$ satisfies weak-mixing condition 2.2.2 with rate $\phi_{\mathcal{C}}(\cdot)$, then for any $T > 0, T \in \mathbb{N}$ stochastic process $Z_t^a := \left(X_t^a \mathbb{I}_{t\leq T}\right)$ also satisfies weak-mixing condition 2.2.2 with the same rate $\phi_{\mathcal{C}}(\cdot)$. Therefore, we can restrict our attention to the processes indexed with the set $\mathbb{N}$. Consider any stochastic process$(X_t)_{t\in\mathbb{N}}$ via its canonical version over $(\Omega, \mathcal{A}, \mathbb{P})$ with $\Omega, \mathcal{A}$ as in 4.3 and measure $\mathbb{P}$ which satisfies Equation (4.4) with rate $\phi_{\mathcal{C}}(\cdot)$. Let $(\mathcal{F}_t)_{t\geq 1}$ be its canonical $\sigma-$algebra generated by a process $\left(\left(X_t^{(a)}\right)_{t\geq 1}, a \in \{K\}\right)$. We consider a sequence of stopping times $(\tau_s)_{s\geq 1}$ with respect to some filtration $(\mathcal{A}_t)_{t\geq 1}$. We consider also the natural filtration of the bandit processes $\mathcal{F}_t = \sigma\{X_u^a, a \in \{K\}, u \leq t\}$, and for every $a \in \{K\}$ we consider $\mathcal{F}_t^a = \sigma(X_s^a, s \leq t)$. For any $s \in \mathbb{N}$, $s \leq s_{end}$ and $\tau_s \in \mathbb{N}$ we denote the following random process $\tilde{X}_t := X_{\tau_s+t}$. Consider the case where $\tau_s = t_0 \in \{T\}$, a stopping time. Consider a regular conditional distribution $\mathbb{P}_{\mathcal{A}_{t_0}}[\cdot] = \mathbb{P}[\cdot|\mathcal{A}_{t_0}]$, and denote the corresponding conditional expectation through $\mathbb{E}_{\mathcal{A}_{t_0}}[\cdot]$. Remember that $\mathbb{P}_{\mathcal{A}_{t_0}} := \mathbb{P}_{\mathcal{A}_{t_0}(\omega)}$ is a random measure, and that we would like to substitute the fixed measure $\mathbb{P}$ with this random measure in Definition 2.2.2.

Notice that the start τ_s of any epoch s is itself random. We show that the mixing property of process $(X_t)_{t\geq 1}$ transfers to the process $\tilde{X}_t$.

Define the following set:

$$\mathcal{A}_{\tau_s} := \{A \in \mathcal{A} : A \cap \{\tau_s \leq t\} \in \mathcal{A}_t, \forall t \in \{T\}\}. \tag{4.13}$$

One can readily check that the set in Equation (4.13) is a $\sigma-$ algebra. Furthermore, we notice that, if τ_s is a stopping time and $s \in \mathbb{N}$ is some fixed number, then we have that $\tau_s + q$ is a stopping time. Define the corresponding σ algebra as:

$$\mathcal{A}_{\tau_s+q} := \{A \in \mathcal{A} : A \cap \{\tau_s + q \leq t\} \in \mathcal{A}_t, \forall t \in \{T\}\}. \tag{4.14}$$

We make use of the following technical measure-theoretic result which prove is folklore and we provide it below for completeness of the narrative.

Lemma 4.7.1. *Let Z be an integrable real-valued random variable, defined over some probability space $(\Omega, \mathcal{A}, \mathbb{P})$ valued in $(\mathcal{X}, \mathcal{B}(\mathcal{X})) \subset (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $\mathcal{A}_{\tau_s}, \mathcal{A}_{\tau_s+q}$ are $\sigma-$algebras as defined in Equations (4.13) and (4.14). Then $\mathbb{P}-$a.s. it holds that for any $\ell \in \{T\}$*

$$\mathbb{I}_{\tau_s=\ell}\mathbb{E}[Z|\mathcal{A}_{\tau_s+q}] = \mathbb{I}_{\tau_s=\ell}\mathbb{E}[Z|\mathcal{A}_{\ell+q}] \tag{4.15}$$

Proof Notice that from the definition of the stopping time τ_s and because $\tau_s \in \mathbb{N}$ we have $\{\tau_s = \ell\} \in \mathcal{A}_\ell \subset \mathcal{A}_{\ell+q}$. It is also straightforward to check that $\{\tau_s = \ell\} \in \mathcal{F}_{\tau_s}$. For any $A \in \mathcal{A}_{\ell+q} \subset \mathcal{A}$ it holds:

$$\begin{aligned}
\mathbb{E}[\mathbb{I}_A\mathbb{I}_{\tau_s=\ell}\mathbb{E}[Z|\mathcal{A}_{\tau_s+q}]] &= \mathbb{E}[\mathbb{I}_{A\cap\tau_s=\ell}\mathbb{E}[Z|\mathcal{A}_{\tau_s+q}]] \\
&= \mathbb{E}[\mathbb{E}[\mathbb{I}_{A\cap\tau_s=\ell}Z|\mathcal{A}_{\tau_s+q}]] \\
&= \mathbb{E}[\mathbb{I}_{A\cap\tau_s=\ell}Z] \\
&= \mathbb{E}[\mathbb{I}_{A\cap\tau_s}\mathbb{E}[Z|\mathcal{A}_{\ell+q}]] \\
&= \mathbb{E}[\mathbb{I}_A\mathbb{I}_{\tau_s=\ell}\mathbb{E}[Z|\mathcal{A}_{\ell+q}]],
\end{aligned}$$

where in the second line we used the fact that $\mathbb{I}_{A\cap\tau_s}$ is $\mathcal{A}_{\tau_s+q}-$measurable, in the third the law of total probability and tower property in the fourth line. The latter equality implies that $\mathbb{I}_A\mathbb{I}_{\tau_s=\ell}\mathbb{E}[Z|\mathcal{A}_{\ell+q}]$ is a

version of $\mathbb{I}_A \mathbb{I}_{\tau_s=\ell} \mathbb{E}[Z|\mathcal{A}_{\tau_s+q}]$ and the claim follows. $\square$

For all $q \in \mathbb{N} \setminus \{0\}$ we have $\mathcal{A}_{\tau_s} \subset \mathcal{A}_{\tau_s+q}$. We denote by $\mathbb{P}_{\mathcal{A}_{\tau_s}} := \mathbb{P}(\cdot|\mathcal{A}_{\tau_s})$ the regular conditional distribution of $\mathbb{P}$ conditioned by the $\sigma-$algebra $\mathcal{A}_{\tau_s}$ and define the conditional mixing coefficients with respect to the random time τ_s as

$$\phi_{\mathcal{C}}^{|\tau_s}(t,\omega) = \sup_{q\geq 1}\Big\{ \big\| \mathbb{E}_{\mathcal{A}_{\tau_s}}[X_{\tau_s+t+q}|\mathcal{A}_{\tau_s+q}] - \mathbb{E}_{\mathcal{A}_{\tau_s}}[X_{\tau_s+t+q}] \big\|_{L_\infty(\mathbb{P})}, \Big\}. \tag{4.16}$$

We show that to control the *random* quantity $\phi_{\mathcal{C}}^{|s}(t,\omega)$ for $\mathbb{P}-$almost all ω it is sufficient to control the following non-random quantity

$$\phi_{\mathcal{C}}^{s}(t) = \sup_{q\geq 1}\{ \|\mathbb{E}[X_{\tau_s+t+q}|\mathcal{A}_{\tau_s+q}] - \mathbb{E}[X_{\tau_s+t+q}|\mathcal{A}_{\tau_s}]\|_{L_\infty(\mathbb{P})}|\}. \tag{4.17}$$

We have the following result for the random time τ_s of the epoch s.

Lemma 4.7.2. *Let $(X_t)_{t\in\mathbb{N}}$ be an arbitrary stochastic process with distribution $\mathbb{P}$ over some measurable space $(\Omega,\mathcal{A})$ and stopping τ_s with respect to filtration $(\mathcal{A}_t)_{t\geq 1}$. Let $\mathcal{A}_{\tau_s}$ be as defined in (4.13); for every $t \in \mathbb{N}$ it holds*

$$\phi_{\mathcal{C}}^{\tau_s(\omega)}(t,\omega) \leq \phi_{\mathcal{C}}^{s}(t).$$

Proof First, notice that for $q \in \mathbb{N}$ and Z_q random variable over $(\Omega,\mathcal{A},\mathbb{P})$ we have $\big\|\sup_{q\in\mathbb{N}} Z_q\big\|_{L_\infty(\mathbb{P})} = \sup_{q\in\mathbb{N}}\|Z_q\|_{L_\infty(\mathbb{P})}$. Thus, applying the latter expression to $Z_q := \mathbb{E}_{\mathcal{A}_{\tau_s}}[X_{\tau_s+t+q}|\mathcal{A}_{\tau_s+q}] - \mathbb{E}_{\mathcal{A}_{\tau_s}}[X_{\tau_s+t+q}]$, we have that in the expression for the conditional mixing coefficient, we can exchange the sup and $\|.\|_\infty$ operations, which implies that

$$\phi_{\mathcal{C}}^{|\tau_s(\omega)}(t,\omega) = \big\| \sup_{q\geq 1}\big\{ \big|\mathbb{E}_{\mathcal{A}_{\tau_s}}[X_{\tau_s+t+q}|\mathcal{A}_{\tau_s+q}] - \mathbb{E}_{\mathcal{A}_{\tau_s}}[X_{\tau_s+t+q}]\big|\big\}\big\|_{L_\infty(\mathbb{P})}. \tag{4.18}$$

Consider arbitrary $C > 0$ and let

$$A\big(\omega,\omega^{'}\big) := \Big\{ \sup_{q\geq 1}\big|\mathbb{E}_{\mathcal{A}_{\tau_s}}[X_{\tau_s+t+q}|\mathcal{A}_{\tau_s+q}] - \mathbb{E}_{\mathcal{A}_{\tau_s}}[X_{\tau_s+t+q}]\big|_{(\omega,\omega')} > C \Big\}, \tag{4.19}$$

which is an event from the probability space $\big(\Omega^2, \mathcal{A}\times\mathcal{A}_{\tau_s}, \mathbb{P}\otimes\mathbb{P}_{\mathcal{A}_{\tau_s}}\big)$. Notice that if $\mathbb{P}-$a.s for every ω, we have that $\mathbb{P}_{\mathcal{A}_{\tau_s}(\omega)}(A) = 0$, then $\mathbb{P}-$a.s. in ω: $\tilde{\phi}_{\mathcal{C}}^{|\tau_s(\cdot)}(t,\cdot) \leq C$. By Fubini's Theorem it is sufficient to show that $\mathbb{P}\otimes\mathbb{P}_{\mathcal{A}_{\tau_s}}\big(A\big(\omega^{'},\omega\big)\big) = 0$. Notice that from the definition of regular conditional distribution we have:

$$\mathbb{P}\otimes\mathbb{P}_{\mathcal{A}_{\tau_s}}(A) = \mathbb{E}_{\mathbb{P}\otimes\mathbb{P}_{\mathcal{A}_{\tau_s}}}\Big[\mathbb{I}_A\big(\omega,\omega^{'}\big)\Big] = \mathbb{E}_{\mathbb{P}}\Big[\mathbb{E}_{\mathbb{P}_{\mathcal{A}_{\tau_s}}}\Big[\mathbb{I}_A\big(\omega,\omega^{'}\big)\Big]\Big] = \mathbb{E}_{\mathbb{P}}[\mathbb{I}_A(\omega,\omega)],$$

where the last quantity follows from the definition of regular conditional distribution $\mathbb{P}_{\mathcal{A}_{\tau_s}(\omega)}$. Finally, to show that $\mathbb{E}_{\mathbb{P}}[\mathbb{I}_A(\omega,\omega)] = 0$ we apply the "diagonal extraction" principle for iterated conditional distributions (see Theorem 6.21 in Kallenberg (2017)). We provide it below for completeness.

Theorem 4.7.3. *[Theorem 6.21 of Kallenberg (2017)] For any probability space $(\Omega,\mathfrak{A},\mathbb{P})$ and Borel-generated $\sigma-$algebras $\mathcal{F},\mathcal{G} \subset \mathcal{A}$, we have that it holds for $(\mathcal{A},\mathbb{P})$-almost all $\omega \in \Omega$:*

$$\mathbb{P}(\cdot|\mathcal{F})(\cdot|\mathcal{G})_{(\omega,\omega)} = \mathbb{P}(\cdot|\mathcal{G})(\cdot|\mathcal{F})_{(\omega,\omega)} = \mathbb{P}(\cdot|\mathcal{F}\vee\mathcal{G})_{(\omega)}, \tag{4.20}$$

where $\mathcal{F}\vee\mathcal{G}$ we define the smallest $\sigma-$ algebra that contains both $\mathcal{F}$ and $\mathcal{G}$.

Applying the result of Theorem 4.7.3 to the σ-algebras $\mathcal{A}_{\tau_s}$ and $\mathcal{A}_{\tau_s+q}$ (note that both of them are

Borel generated by the canonical version of $(X_t)_{t\in\mathbb{N}}$ over $(\Omega, \mathcal{A}_{\tau_s})$ and $(\Omega, \mathcal{A}_{\tau_s+q})$ correspondingly), and noticing that $\mathcal{A}_{\tau_s} \subset \mathcal{A}_{\tau_s+s}$) we deduce that $\mathbb{P}$-a.s.:

$$\mathbb{P}(\cdot|\mathcal{A}_{\tau_s})(\cdot|\mathcal{A}_{\tau_s+q})_{(\omega,\omega)} = \mathbb{P}(\cdot|\mathcal{A}_{\tau_s+q})_{(\omega)},$$

therefore

$$\mathbb{E}_{\mathcal{A}_{\tau_s}}[X_{\tau_s+t+s}|\mathcal{A}_{\tau_s+s}]_{(\omega,\omega)} = \mathbb{E}[X_{\tau_s+t+s}|\mathcal{A}_{\tau_s+s}]_{(\omega)}.$$

The latter implies that $\mathbb{P}-$ a.s. we have

$$A(\omega,\omega) = \{\sup_{q\geq 1}|\mathbb{E}[X_{\tau_s+t+q}|\mathcal{A}_{\tau_s+q}] - \mathbb{E}[X_{\tau_s+t+q}|\mathcal{A}_{\tau_s}]| > C\},$$

and the claim of the Lemma follows by taking $C = \widetilde{\phi}^s_{\mathcal{C}}(t)$. □

Theorem 4.7.4. *Let τ_s be the stopping time from the Equation (4.13), such that $\mathbb{P}[\tau_s < \infty] = 1$. For all $t, q > 0$ it holds that*

$$\|\mathbb{E}[X_{\tau_s+t+q}|\mathcal{A}_{\tau_s+q}] - \mathbb{E}[X_{\tau_s+t+q}]\|_{L_\infty(\mathbb{P})} \leq \phi_{\mathcal{C}}(t).$$

Proof Denote $Y_{\tau_s+t+q} := X_{\tau_s+t+q} - \mathbb{E}[X_{\tau_s+t+q}]$. Obviously Y_{τ_s+t+q} is measurable w.r.t. the $\sigma-$algebra $\mathcal{A}_{\tau_s+t+q}$ and is a centered random variable. Because $\mathbb{I}_{\tau_s=\ell}$ is $\mathcal{A}_{\tau_s}-$measurable for any $\ell \in \{T\}$ and $\mathcal{A}_{\tau_s} \subset \mathcal{A}_{\tau_s+q}$ for any $q \geq 1$, we can write

$$\mathbb{E}[Y_{\tau_s+q+t}|\mathcal{A}_{\tau_s+q}] = \mathbb{E}\left[\sum_{\ell\in T}\mathbb{I}_{\tau_s=\ell}Y_{\tau_s+q+t}|\mathcal{A}_{\tau_s+q}\right] = \sum_{\ell\in T}\mathbb{I}_{\tau_s=\ell}\mathbb{E}[Y_{\ell+q+t}|\mathcal{A}_{\tau_s+q}].$$

Applying Lemma 4.7.1 to $Z := Y_{\tau_s+q+t} = X_{\tau_s+q+t} - \mathbb{E}[X_{\tau_s+q+t}]$, which is real-valued and defined on the space $(\Omega, \mathcal{A}, \mathbb{P})$ we have that it holds $\mathbb{P}-$a.s :

$$\begin{aligned}\mathbb{E}[Y_{\tau_s+q+t}|\mathcal{A}_{\tau_s+q}] &= \sum_{\ell\in\{T\}}\mathbb{E}[\mathbb{I}_{\tau_s=\ell}Y_{\tau_s+q+t}|\mathcal{A}_{\tau_s+q}] = \sum_{\ell\in\{T\}}\mathbb{I}_{\tau_s=\ell}\mathbb{E}[Y_{\ell+q+t}|\mathcal{A}_{\tau_s+q}] \\ &= \sum_{\ell\in\{T\}}\mathbb{I}_{\tau_s=\ell}\mathbb{E}[Y_{\ell+q+t}|\mathcal{A}_{\ell+q}] \leq \sum_{\ell\in\{T\}}\mathbb{I}_{\tau_s=\ell}\phi_{\mathcal{C}}(t) = \phi_{\mathcal{C}}(t).\end{aligned}$$

Therefore, we finally get:

$$\left\|\mathbb{E}_{\mathcal{A}_{\tau_s}}\left[\tilde{X}_{t+q}\right] - \mathbb{E}\left[\tilde{X}_{t+q}\right]\right\|_{L_\infty(\mathbb{P})} = \|\mathbb{E}[Y_{\tau_s+q+t}|\mathcal{A}_{\tau_s+q}]\|_{L_\infty(\mathbb{P})} \leq \phi_{\mathcal{C}}(t),$$

and the claim is proved. □

Now we have all the ingredients to prove Proposition 4.3.1.

Proof of Proposition 4.3.1 Because the bandit instance $\mathbb{P}$ is weakly-mixing with rate $\phi_{\mathcal{C}}(\cdot)$ and **Assumption 1** holds, we have that every process $(X^a_t)_{t\geq 1}$ is $\mathcal{C}-$weakly mixing with rate $\phi_{\mathcal{C}}(\cdot)$. For a random time τ_s by Lemma 4.7.2 it holds that $\tilde{\phi}^{\tau_s(\omega)}_{\mathcal{C}}(t,\omega) \leq \tilde{\phi}^s_{\mathcal{C}}(t)$. Thus, from the bound of Lemma (4.7.2), triangle inequality, the result of Theorem 4.7.4 we have

$$\begin{aligned}&\left\|\mathbb{E}\left[\tilde{X}_{t+q}|\mathcal{A}_{\tau_s+q}\right] - \mathbb{E}\left[\tilde{X}_{t+q}|\mathcal{A}_{\tau_s}\right]\right\|_{L_\infty(\mathbb{P})} \\ &\leq \left\|\mathbb{E}\left[\tilde{X}_{t+q}|\mathcal{A}_{\tau_s+q}\right] - \mathbb{E}\left[\tilde{X}_{t+q}\right]\right\|_{L_\infty(\mathbb{P})} + \left\|\mathbb{E}\left[\tilde{X}_{t+q}|\mathcal{A}_{\tau_s}\right] - \mathbb{E}\left[\tilde{X}_{t+q}\right]\right\|_{L_\infty(\mathbb{P})} \\ &\leq 2\phi_{\mathcal{C}}(t).\end{aligned}$$

Taking the supremum over all $q \geq 1$, we obtain $\phi_{\mathcal{C}}{}^{s}(t) \leq 2\phi_{\mathcal{C}}(t)$, which, together with the bound $\phi_{\mathcal{C}}^{\tau_s(\omega)}(t,\omega) \leq \phi_{\mathcal{C}}^{s}(t)$ from Lemma 4.7.2, implies the claim of the Proposition. □

4.7.2 Proof of Theorem 4.3.8

First, we prove the intermediate result, which ensures the optimal choice of the number of pulls in the slow–mixing scenario for the given epoch s in the learning scheme of Algorithm 2.

Lemma 4.7.5. *Consider any epoch $s \in \mathbb{N}$; let θ_s, δ_s and $\Omega(\theta_s, b_s)$ be chosen as in Algorithm 2. Define B_s as the set of active arms at the epoch s and ς_s the corresponding pulling strategy. We assume that the bandit instance is $\mathcal{C}-$weakly mixing with the rate $\phi_{\mathcal{C}}(t) = t^{-\alpha}$, and $\alpha \in (0, 1/2)$. Furthermore, denote $c_0 = ((1-\alpha)(1/2-\alpha))^{-1}$, $c_1 = \left(\frac{c_0}{80}\right)^{\frac{2}{1-2\alpha}}$, $c_3 = 12800c_0$ and s_{end} for the last possible epoch. If the number of pulls T_s of each arm $j \in B_s$ at the epoch s is chosen as*

$$T_s = T_{s,1} := \left\lceil \frac{32\log\left(\mathcal{A}T\theta_s^2\right)}{\theta_s^2} \right\rceil, \text{ for } b_s \geq \left(32c_1^{-1}\theta_s^{-2}\log\left(\mathcal{A}T\theta_s^2\right)\right)^{\frac{1-2\alpha}{2\alpha}};$$

$$T_s = T_{s,2} := \left\lceil \frac{1}{b_s}\left(\frac{c_3\log\left(\mathcal{A}T\theta_s^2\right)}{\theta_s^2}\right)^{\frac{1}{2\alpha}} \right\rceil, \text{ for } b_s \leq \left(32c_1^{-1}\theta_s^{-2}\log\left(\mathcal{A}T\theta_s^2\right)\right)^{\frac{1-2\alpha}{2\alpha}};$$

then it holds that $\Omega(\theta_s, b_s) \leq \frac{\theta_s}{2}$ for $s \in \{0, \dots, s_{end}\}$.

Proof Without loss of generality, we enumerate all arms in B_s as $\{1, \dots, b_s\}$. For an arm $j \in B_s$ we pull it according to the equispaced schedule ς_s^j, defined as follows:

$$\varsigma_s^j = (j, j + b_s, \dots, j + (T_s - 1)b_s).$$

The total number of pulls of every arm j during the epoch is $\left|\varsigma_s^j\right| = T_s$ for each $j \in \{1, 2, \dots, b_s\}$. Notice that for $0 < \alpha < \frac{1}{2}$ then we have

$$\sum_{j=1}^{T_s} j^{-\frac{3}{2}} \sum_{\ell=1}^{j} \phi_{\mathcal{C}}(b_s\ell) = b_s^{-\alpha} \sum_{j=1}^{T_s} j^{-\frac{3}{2}} \sum_{\ell=1}^{j} \ell^{-\alpha} \leq c_0 b_s^{-\alpha} T_s^{\frac{1}{2}-\alpha},$$

with $c_0 = ((1-\alpha)(1/2-\alpha))^{-1}$.

Thus, plugging this inequality into the confidence term $\Omega(\theta_s, b_s)$ from Algorithm 2 in the epoch s, we deduce:

$$\Omega(\theta_s, b_s) \leq 2\max\left(1, 80c_0 b_s^{-\alpha} T_s^{1/2-\alpha}\right)\sqrt{\frac{2\log(\mathcal{A}T\theta_s^2)}{T_s}}. \tag{4.21}$$

Now, if $b_s > \left(32c_1^{-1}\theta_s^{-2}\log\left(\mathcal{A}T\theta_s^2\right)\right)^{\frac{1-2\alpha}{2\alpha}}$, then $80c_0 b_s^{-\alpha} T_s^{1/2-\alpha} \leq 1$, so if we choose

$$T_s := \left\lceil \frac{32\log\left(\mathcal{A}T\theta_s^2\right)}{\theta_s^2} \right\rceil$$

and plug into Equation (4.21), we ensure that $\Omega(\theta_s, b_s) \leq \frac{\theta}{2}$.

Similarly, if $b_s \leq \left(32c_1^{-1}\theta_s^{-2}\log\left(\mathcal{A}T\theta_s^2\right)\right)^{\frac{1-2\alpha}{2\alpha}}$ we have that $80c_0 b_s^{-\alpha}T_s^{1/2-\alpha} > 1$, so by choosing

$$T_s := \left\lceil \frac{1}{b_s}\left(\frac{12800\log\left(\mathcal{A}T\theta_s^2\right)}{\theta_s^2}\right)^{\frac{1}{2\alpha}}\right\rceil,$$

we assure also that $\Omega(\theta_s, b_s) \leq \frac{\theta}{2}$. Therefore, the lemma is proved. □

Combining the choice of the number of pulls T_s given in Lemma 4.7.5, the theoretical argument that the $\phi_{\mathcal{C}}-$mixing property of the processes $(X_{\tau_s+t})_{t\in\{T\}}$ is preserved under time-transform (provided that τ_s is a stopping time), and making use of concentration bounds (4.6) for mixing samples collected during each epoch, we now can prove the main result.

Proof of the Theorem 4.3.8. Recall that we denote for $\lambda \geq 0$ the set A_λ as the set of suboptimal arms i for which $\{\Delta_i > \lambda\}$ so that A_0 is the overall set of suboptimal arms. Recall that $\Delta_k := \mu_* - \mu_k$ for $k \in A_0$ and we define $\Delta_{*,\lambda} = \min_{j\in A_\lambda}\Delta_j$. For an epoch s consider confidence bound $\Omega(\theta_s, b_s)$ as in Algorithm 2, where b_s is the number of active arms during the epoch s. b_s is random quantity, which is $\mathcal{F}^I_{\tau_s}-$measurable for every s, $\theta_s = 2^{-s}$. For the regret, we keep the dependence on the $\mathcal{C}-$Mix UCB Algorithm implicit and write $R_{\mathbb{P}}(\mathcal{C}_{mix}, T) := R_{\mathbb{P}}(T) = \mathbb{E}\left[\sum_{k\in A_0}\Delta_k N_k(T)\right] = \sum_{k\in A_0}\Delta_k\mathbb{E}[N_k(T)]$; the main target is to upper bound the number of pulls of each arm $k \in A_0$. We suppress index T in $N_k(T)$ for simplicity. For every suboptimal arm i, define $m_i := \min\{m \in \mathbb{N} : \theta_m \leq \frac{\Delta_i}{2}\}$. From the definitions of m_i and of θ_i it follows that $\frac{\Delta_i}{4} \leq \theta_{m_i} < \frac{\Delta_i}{2} \leq \theta_{m_i-1}$. Solving this as the inequality in $\frac{1}{\Delta_i}$, we get:

$$\frac{1}{\theta_{m_i}} \leq \frac{4}{\Delta_i} < \frac{1}{\theta_{m_i+1}}. \tag{4.22}$$

We fix some best arm (which we later refer to as $*$) and denote by M_* the first epoch when this optimal arm $*$ has been eliminated. Note that it is possible that $M_* = \infty$, and that it is enough to consider only a certain optimal arm for the further analysis. Also, let $m_\lambda := \min\{m|\theta_m < \frac{\lambda}{2}\}$, which implies that for all $i \in A_\lambda$ we have $m_i \leq m_\lambda$.

For the arm $k \in A_\lambda$, let M_k be the (random) epoch at which arm k is eliminated. Consider the following event:

$$\mathcal{E}_k = \{\tau_{m_k+1} \leq T, M_k \leq m_k\} \cup \{\tau_{m_k+1} > T\}.$$

Notice that $\mathcal{E}_k^c = \{\tau_{m_k+1} \leq T, M_k > m_k\}$, which means that the arm k is eliminated after epoch m_k, and m_k is finished. Using the definition of event $\mathcal{E}_k$ for each $k \in \mathcal{E}_k$, and introducing an arbitrary threshold $\lambda > 0$, we decompose the pseudo-regret into the following parts:

$$\begin{aligned}
R_{\mathbb{P}}(T) &= \mathbb{E}\left[\sum_{k\in A_0}\Delta_k N_k\right] = \sum_{k\in A_0}\mathbb{E}[N_k]\Delta_k \\
&\leq \sum_{k\in A_\lambda}\mathbb{E}[N_k]\Delta_k + \max_{k\in A_0\setminus A_\lambda}\Delta_k\mathbb{E}\left[\sum_{k\in A_0\setminus A_\lambda}N_k\right] \\
&= \sum_{k\in A_\lambda}\mathbb{E}[N_k\mathbf{1}\{\mathcal{E}_k\}]\Delta_k + \sum_{k\in A_\lambda}\mathbb{E}[N_k\mathbf{1}\{\mathcal{E}_k^c\}]\Delta_k + \max_{k:A_0\setminus A_\lambda}\Delta_k\mathbb{E}\left[\sum_{k\in A_0\setminus A_\lambda}N_k\right].
\end{aligned}$$

We analyze the contributions from each of the sums in the last inequality separately. Clearly, the last sum can be bounded by $T\max_{k:A_0\setminus A_\lambda}\Delta_k$.

For every round k we consider whether the optimal arm is being eliminated before or after round of "high probability elimination" of arm k (i.e., at m_k). We have that first sum can be decomposed in the following way:

$$\sum_{k\in A_\lambda} \Delta_k \mathbb{E}[N_k \mathbf{1}\{\mathcal{E}_k^c\}] = \sum_{k\in A_\lambda} \Delta_k \mathbb{E}[N_k \mathbf{1}\{\mathcal{E}_k^c\}\mathbf{1}\{M_* < m_k\}] + \sum_{k\in A_\lambda} \Delta_k \mathbb{E}[N_k \mathbf{1}\{\mathcal{E}_k^c\}\mathbf{1}\{M_* \geq m_k\}]. \tag{4.23}$$

Consider the second sum on the right hand side in Equation (4.23). For a fixed arm $k \in A_\lambda$, the confidence level of the epoch s is selected as $\delta_s = \frac{1}{T\theta_s^2}$. For each $k \in \{K\}$, $s \in \{0, \dots, s_{\text{end}}\}$ we consider events $D_{k,s}$ and $E_{*,s}$, whose complements are given as:

$$D_{k,s}^c := \{k \in B_s, \tau_{s+1} \leq T, \widehat{\mu}_{k,s} \leq \mu_k + \Omega(\theta_s, b_s)\},$$
$$E_{*,s}^c := \{* \in B_s, \tau_{s+1} \leq T, \widehat{\mu}_{*,s} \geq \mu_* - \Omega(\theta_s, b_s)\},$$

where $\Omega(\theta_s, b_s)$ is as in Algorithm 2. We remark that conditions $\{\tau_{s+1} \leq T\}$ and either $\{k \in B_s\}$ or $\{* \in B_s\}$ are added to assure that the computation procedure of Algorithm 2 can be formally completed in the epoch s. By Proposition 4.3.1, the ordered set of samples that are collected during epoch s from the process $X^k_{\tau_s+t}$, $k \in B_s$ satisfy, conditionally to the $\sigma-$algebra $\mathcal{F}^I_{\tau_s}$, a weak-mixing assumption with rate $2\phi_{\mathcal{C}}(t)$. Thus, we can apply (conditioned on the information given at the beginning of the epoch) concentration inequality (4.6) for the weak-mixing process $X^k_{\tau_s+t}$ to control the probabilities of the "bad" events $D^c_{k,s}, E^c_{*,s}$. The latter implies that we have that $\mathbb{P}-$almost surely holds:

$$\mathbb{P}_{\mathcal{F}_{\tau_s}}\left[D^c_{k,s}\right] \leq \delta_s, \qquad \mathbb{P}_{\mathcal{F}_{\tau_s}}\left[E^c_{*,s}\right] \leq \delta_s. \tag{4.24}$$

For the moment, we are interested in the case where $s = m_k$. Notice that the event $\mathbf{1}\{\mathcal{E}_k^c\}\mathbf{1}\{M_* \geq m_k\}$ implies that arm k has not been eliminated until the epoch m_k, while arm $*$ belongs to the set of active arms B_{m_k} and the epoch m_k has been completed. Furthermore, one can readily check that event $\mathcal{E}_k^c \cap \{M_* \geq m_k\}$ implies that the event $D^c_{k,m_k} \cup E^c_{*,s}$ holds. To prove this, notice that on the event $\mathcal{E}_k^c \cap \{M_* \geq m_k\}$ it holds that $\tau_{m_k+1} \leq T$, $* \in B_{m_k}$ and $k \in B_{m_k}$. Now, if neither $\{\widehat{\mu}_{k,m_k} > \mu_k + \Omega(\theta_s, b_s)\}$ nor $\{\widehat{\mu}_{*,m_k} > \mu_* + \Omega(\theta_s, b_s)\}$ holds, then arm k will be eliminated at the end of epoch m_k. Indeed, by using Lemma 4.7.5 and Inequality (4.22), we have that $\Omega(\theta_{m_k}, b_{m_k}) \leq \frac{\theta_{m_k}}{2} \leq \frac{\Delta_k}{4}$. This implies the following chain of inequalities:

$$\begin{aligned}\widehat{\mu}_{k,s} + \Omega(\theta_{m_k}, k_{m_k}) &\leq \mu_k + 2\Omega(\theta_{m_k}, b_{m_k}) \\ &\leq \mu_k + \Delta_k - 2\Omega(\theta_{m_k}, b_{m_k}) \\ &= \mu_* - 2\Omega(\theta_{m_k}, b_{m_k}) \leq \widehat{\mu}_{*,s} - \Omega(\theta_{m_k}, b_{m_k}),\end{aligned}$$

and the arm k is eliminated because of the scheme of Algorithm 2. Therefore, we have $\mathcal{E}_k^c \cap \{M_* \geq m_k\} \subset D^c_{k,m_k} \cup E^c_{*,m_k}$. Furthermore, notice that by conditioning on the $\sigma-$algebra $\mathcal{F}_{\tau_{m_k}}$ of the events preceding the epoch m_k, we get:

$$\mathbb{E}[\mathbf{1}\{\mathcal{E}_k^c\}\mathbf{1}\{M_* \geq m_k\}] = \mathbb{E}\Big[\mathbb{E}_{\mathcal{F}_{\tau_{m_k}}}[\mathbf{1}\{\mathcal{E}_k^c\}\mathbf{1}\{M_* \geq m_k\}]\Big] \leq \mathbb{E}\Big[\mathbb{P}_{\mathcal{F}_{\tau_{m_k}}}\left[D^c_{k,m_k} \cup E^c_{*,m_k}\right]\Big] \leq \frac{2}{\mathcal{A}T\theta^2_{m_k}},$$

where in the last inequality we used the union bound over events E^c_{*,m_k}, D^c_{k,m_k}, their control by means of (conditional) concentration inequality for the samples collected during the epoch m_k. Thus, bounding the number of pulls N_k trivially by T and plugging in the bound on the $\mathbb{E}[\mathbf{1}\{\mathcal{E}_k^c\}\mathbf{1}\{M_* \geq m_k\}]$, we

obtain for this part of the pseudo-regret the following upper bound:

$$\sum_{k\in A_\lambda}\Delta_k\mathbb{E}[N_k\mathbf{1}\{\mathcal{E}_k^c\}\mathbf{1}\{M_*\geq m_k\}]\leq\sum_{k\in A_\lambda}\Delta_k T\frac{2}{\mathcal{A}T\theta_{m_k}^2}\leq\frac{32}{\mathcal{A}}\sum_{k\in A_\lambda}\frac{1}{\Delta_k},$$

where in the last inequality we used the relation (4.22) between θ_{m_k} and Δ_{m_k}. We focus now on the first sum term in the right hand side of Equation (4.23). By changing the order of summation over the epochs and counting the regret from the active arms in each epoch s (which is at most $2\theta_s$ for every arm) and upper bounding the number of rounds by T, we obtain:

$$\begin{aligned}
\sum_{k\in A_\lambda}\Delta_k\mathbb{E}[N_k\mathbf{1}\{\mathcal{E}_k^c\}\mathbf{1}\{M_*<m_k\}]&\leq\sum_{k\in A_\lambda}\Delta_k\sum_{s<m_k}\mathbb{E}[N_k\mathbf{1}\{M_*=s\}\mathbf{1}\{\tau_{m_k+1}\leq T,M_k>m_k\}]\\
&=\sum_{s=0}^{m_\lambda}\sum_{k:m_k>s}\Delta_k\mathbb{E}[N_k\mathbf{1}\{M_*=s\}\mathbf{1}\{\tau_{m_k+1}\leq T,M_k>m_k\}]\\
&\leq\sum_{s=0}^{m_\lambda}2\theta_s\mathbb{E}\left[\mathbf{1}\{M_*=s\}\mathbf{1}\{\tau_{s+1}\leq T\}\sum_{k:m_k>s}N_k\right]\\
&\leq 2\sum_{s=0}^{m_\lambda}T\theta_s\mathbb{P}[M_*=s,\tau_{s+1}\leq T].
\end{aligned}$$

Recall that through B_s we denote the set of active arms at the epoch s and b_s is its cardinality. Because the event $M_*=s$ means that the optimal arm was eliminated by some active arm k in the epoch s, we have

$$\begin{aligned}
\mathbb{P}[M_*=s,\tau_{s+1}\leq T]&\leq\mathbb{E}\left[\mathbf{1}\{*\in B_s,\tau_{s+1}\leq T\}\sum_{k\in B_s}\mathbf{1}\{\widehat{\mu}_{k,s}>\widehat{\mu}_{*,s}+2\Omega(\theta_s,b_s)\}\right]\\
&\leq\sum_{k:m_k\geq s}\mathbb{E}\big[\mathbf{1}\{k\in B_s;*\in B_s;\tau_{s+1}\leq T,\widehat{\mu}_{k,s}>\widehat{\mu}_{*,s}+2\Omega(\theta_s,b_s)\}\big]\\
&+\mathbb{E}\left[\sum_{k:m_k<s}\mathbf{1}\{*\in B_s;k\in B_s,\tau_{s+1}\leq T\}\right]\\
&\leq\sum_{k:m_k\geq s}\mathbb{E}\big[\mathbb{E}_{\mathcal{F}_{\tau_s}}\big[\mathbf{1}\{k\in B_s;*\in B_s;\tau_{s+1}\leq T,\widehat{\mu}_{k,s}>\widehat{\mu}_{*,s}+2\Omega(\theta_s,b_s)\}\big]\big]\\
&+\sum_{k:m_k<s}\mathbb{E}\Big[\mathbb{E}_{\mathcal{F}_{\tau_{m_k}}}[\mathbf{1}\{*\in B_s;k\in B_s,\tau_{s+1}\leq T\}]\Big]\\
&\leq\sum_{k:m_k\geq s}\frac{2}{\mathcal{A}T\theta_s^2}+\sum_{k:m_k<s}\mathbb{E}\Big[\mathbb{P}_{\mathcal{F}_{\tau_{m_k}}}[\mathcal{E}_k^c\cap\{M_*\geq m_k\}]\Big]\\
&\leq\sum_{k:m_k\geq s}\frac{2}{\mathcal{A}T\theta_s^2}+\sum_{k:m_k<s}\frac{2}{\mathcal{A}T\theta_{m_k}^2},
\end{aligned}$$

where we used tower property for expectations, and that for conditional probabilities it holds almost surely

$$\mathbb{P}_{\mathcal{F}_{\tau_s}}\big[k\in B_s;*\in B_s;\widehat{\mu}_{k,s}>\widehat{\mu}_{*,s}+2\Omega(\delta_s,\varsigma_s)\big]\leq\mathbb{P}_{\mathcal{F}_{\tau_s}}\big[D_{k,s}^c\cup E_{*,s}^c\big]$$

as well as

$$\mathbb{P}_{\mathcal{F}_{\tau_{m_k}}}[\mathcal{E}_k^c \cap \{M_* \geq m_k\}] \leq \mathbb{P}_{\mathcal{F}_{\tau_{m_k}}}[\mathcal{E}_k^c] \leq \mathbb{P}_{\mathcal{F}_{\tau_{m_k}}}\left[D_{k,m_k}^c \cup E_{*,m_k}^c\right]$$

and the control of the event's probabilities in the epoch s given by Equation (4.24). Plugging this bound into the previous result and using the definition of the sequence θ_s, we obtain the following upper bound:

$$\begin{aligned}
2\sum_{s=0}^{m_\lambda} T\theta_s \mathbb{P}[M_* = s, \tau_{s+1} \leq T] &\leq \frac{4}{\mathcal{A}} \sum_{s=0}^{m_\lambda} \theta_s \left(\sum_{k:m_k \geq s} \frac{1}{\theta_s^2} + \sum_{k:m_k < s} \frac{1}{\theta_{m_k}^2} \right) \\
&\leq \frac{4}{\mathcal{A}} \sum_{k \in A_0} \left(\sum_{s \leq m_k \wedge m_\lambda} \frac{1}{\theta_s} + \frac{1}{\theta_{m_k}^2} \sum_{s=m_k+1}^{m_\lambda} \theta_s \right) \\
&\leq \frac{8}{\mathcal{A}} \sum_{k \in A_0} \left(\frac{1}{\theta_{m_k \wedge m_\lambda}} + \frac{\mathbf{1}\{m_k \leq m_\lambda\}}{\theta_{m_k}} \right) \\
&\leq \frac{64}{\mathcal{A}} \left(\sum_{k \in A_\lambda} \frac{1}{\Delta_k} + \sum_{k \in A_0 \setminus A_\lambda} \frac{1}{\lambda} \right),
\end{aligned}$$

Gathering upper bounds for each sum in Equation (4.23), we obtain:

$$\sum_{k \in A_\lambda} \Delta_k \mathbb{E}[N_k \mathbf{1}\{\mathcal{E}_k^c\}] \leq \frac{96}{\mathcal{A}} \sum_{k \in A_\lambda} \frac{1}{\Delta_k} + \frac{64}{\mathcal{A}} \sum_{k \in A_0 \setminus A_\lambda} \frac{1}{\lambda}. \tag{4.25}$$

Finally, for the contribution of $\mathbb{E}\left[\sum_{k \in A_\lambda} N_k \mathbf{1}\{\mathcal{E}_k\}\Delta_k\right]$, we provide the arguments for the quantity under expectation which holds $\mathbb{P}-$ a.s. First, notice that on the event $\mathbf{1}\{\mathcal{E}_k\}$, each arm is pulled until it is eliminated at the latest at the epoch m_k. Thus, recalling that for any $i \in A_\lambda$ $\Delta_i \leq \lambda$ and using simply $T_s \leq T_{s,1} + T_{s,2}$ (where $T_{s,1}, T_{s,2}$ are given by Lemma 4.7.5), we can write the following chain of inequalities, which hold almost surely:

$$\begin{aligned}
\sum_{k \in A_\lambda} N_k \mathbf{1}\{\mathcal{E}_k\}\Delta_k &= \sum_{k \in A_\lambda} \sum_{s=0}^{s_{end} \wedge m_\lambda} \Delta_k T_s \mathbf{1}\{\mathcal{E}_k\}\mathbf{1}\{k \in B_s\} \leq \sum_{k \in A_\lambda} \sum_{s=0}^{s_{end} \wedge m_\lambda} \Delta_k (T_{s,1} + T_{s,2}) \mathbf{1}\{\mathcal{E}_k\}\mathbf{1}\{k \in B_s\} \\
&\leq \sum_{k \in A_\lambda} \Delta_k \sum_{s=0}^{s_{\text{end}} \wedge m_\lambda} T_{s,1} \mathbf{1}\{\mathcal{E}_k\}\mathbf{1}\{k \in B_s\} + \sum_{s=0}^{s_{end} \wedge m_\lambda} 2\theta_s T_{s,2} \sum_{k \in A_\lambda} \mathbf{1}\{k \in B_s\}\mathbf{1}\{\mathcal{E}_k\}
\end{aligned}$$

where in the second sum we exchanged the sums over the arms and over the contribution of each epoch, used the fact that for the arm k active in the epoch s we have that $s \leq m_k$ and thus we pay a regret of order at most $2\theta_s$ by pulling this arm. For the second sum, we observe that the sequence $(\theta_s)^{1-\frac{1}{\alpha}} \left(\log\left(\mathcal{A}T\theta_s^2\right)\right)^{\frac{1}{2\alpha}}$ is monotonically increasing for all $s \leq s_{\text{end}}$ with ratio at least $\frac{6}{5}\left(\sqrt{\frac{12}{5}}\right)^{\frac{1}{\alpha}-2}$ and that

$$\sum_{k \in A_\lambda} \mathbf{1}\{k \in B_s\}\mathbf{1}\{\mathcal{E}_k\} \leq b_s.$$

Therefore, we can write:

$$\begin{aligned}\sum_{s=0}^{s_{\text{end}}\wedge m_\lambda} 2\theta_s T_{s,2}\sum_{k\in A_\lambda}\mathbf{1}\{k\in B_s\}\mathbf{1}\{\mathcal{E}_k\} &\le 4\sum_{s=0}^{s_{\text{end}}\wedge m_\lambda}\theta_s\left(\frac{c_3\log\left(\mathcal{A}T\theta_s^2\right)}{\theta_s^2}\right)^{\frac{1}{2\alpha}}\frac{1}{b_s}\sum_{k\in A_\lambda}\mathbf{1}\{k\in B_s\}\mathbf{1}\{\mathcal{E}_k\}\\ &\le 4c_3^{\frac{1}{2\alpha}}\sum_{s=0}^{s_{\text{end}}\wedge m_\lambda}\theta_s^{1-\frac{1}{\alpha}}\left(\log\left(\mathcal{A}T\theta_s^2\right)\right)^{\frac{1}{2\alpha}}\\ &\le 4c_3^{\frac{1}{2\alpha}}c_4\theta_{s_{\text{end}}\wedge m_\lambda+1}^{1-\frac{1}{\alpha}}\left(\log\left(\mathcal{A}T\theta^2_{s_{\text{end}}\wedge m_\lambda+1}\right)\right)\\ &\le 2^{\frac{1}{\alpha}+1}c_4c_3^{\frac{1}{2\alpha}}\left(\frac{\Delta_{*,\lambda}}{4}\right)^{1-\frac{1}{\alpha}}\left(\log\left(\frac{\mathcal{A}T}{4}\Delta_{*,\lambda}^2\right)\right),\end{aligned}$$

where we used c_3 as in Lemma 4.7.5 set $c_4 = \frac{1}{1.2*\sqrt{2.4}^{\frac{1}{\alpha}-2}-1}$ and used the definition of $T_{s,2}$ from Algorithm 2, and that the contribution of the regret of active arm $k\in A_\lambda$ is at most $\theta_{m_\lambda}\le\frac{\Delta_{*,\lambda}}{4}$ at the end. For the first sum, by plugging in the expression for $T_{s,1}\ge 1$ and using the assumption that in epoch s we sum up the contributions of the regret of the arms k which have not been eliminated until round m_k, we get:

$$\begin{aligned}\sum_{k\in A_\lambda}\Delta_k\sum_{s=0}^{s_{\text{end}}\wedge m_\lambda}T_{s,1}\mathbf{1}\{\mathcal{E}_k\}\mathbf{1}\{k\in B_s\} &\le 2\sum_{k\in A_\lambda}\Delta_k\sum_{s=0}^{s_{\text{end}}\wedge m_\lambda}\frac{32\log\left(\mathcal{A}T\theta_s^2\right)}{\theta_s^2}\mathbf{1}\{\mathcal{E}_k\}\mathbf{1}\{k\in B_s\}\\ &\le 64\sum_{k\in A_\lambda}\Delta_k\sum_{s=0}^{s_{\text{end}}\wedge m_\lambda}\theta_s^{-2}\log\left(AT\theta_s^2\right)\mathbf{1}\{\mathcal{E}_k\}\mathbf{1}\{k\in B_s\}\\ &\le 256\sum_{k\in A_\lambda}\Delta_k\theta_{m_k}^{-2}\log\left(\mathcal{A}T\theta_{m_k}^2\right)\le 1024\sum_{k\in A_\lambda}\Delta_k^{-1}\log\left(\mathcal{A}T\Delta_k^2\right),\end{aligned}$$

where in the last line we used the geometrical increase of the series $\theta_s\log\left(\mathcal{A}T\theta_s^2\right)$ and the relation (4.22) between θ_{m_k} and Δ_k.

Summing up all the terms, we have the following upper bound:

$$\sum_{k\in A_\lambda}\mathbb{E}[N_k\mathbf{1}\{\mathcal{E}_k\}]\Delta_k\le\sum_{k\in A_\lambda}1024\Delta_k^{-1}\log\left(\mathcal{A}T\Delta_k^2\right)+2^{-\frac{1}{\alpha}+3}c_4c_3^{\frac{1}{2\alpha}}(\Delta_{*,\lambda})^{1-\frac{1}{\alpha}}\left(\log\left(\frac{\mathcal{A}T\Delta_{*,\lambda}^2}{4}\right)\right)^{\frac{1}{2\alpha}}. \tag{4.26}$$

Summing up the individual contributions of inequalities (4.26) and (4.25), we obtain

$$\begin{aligned}R_{\mathbb{P}}(T)\le 2\sum_{k\in A_\lambda}&\left(\Delta_k+\frac{96}{\mathcal{A}}\frac{1}{\Delta_k}+512\Delta_k^{-1}\log\left(\mathcal{A}T\Delta_k^2\right)\right)\\ &+\underbrace{2^{-\frac{1}{\alpha}+3}c_4c_3^{\frac{1}{2\alpha}}}_{\tilde{c}}(\Delta_{*,\lambda})^{1-\frac{1}{\alpha}}\left(\log\left(\frac{\mathcal{A}T\Delta_{*,\lambda}^2}{4}\right)\right)^{\frac{1}{2\alpha}}+\frac{64}{\mathcal{A}}\sum_{k\in A_0\setminus A_\lambda}\frac{1}{\lambda}+T\max_{k:A_0\setminus A_\lambda}\Delta_k.\end{aligned}$$

Finally, by noticing that $\Delta_k\le 1$, we imply the statement of the Theorem. □

4.7.3 Proof of Theorem 4.3.12

Proof We give the main argument while using constants C_1, C_2 relatively arbitrary. First, for any choice $\lambda > \frac{1}{2}\sqrt{\frac{e^{1/2}}{T}}$, we have that $C_1\sum_{k\in A_\lambda}\frac{\log\left(\mathcal{A}T\Delta_k^2\right)}{\Delta_k}+\sum_{k\in A_0\setminus A_\lambda}\frac{1}{\lambda}\le\tilde{C}\frac{K\log T}{\lambda}$ (where constants C_1 and $\tilde{C}$

are some numerical constants which are not further precised). Furthermore, from the definition of $\Delta_{*,\lambda}$, because $1 - \frac{1}{\alpha} < 0$ we have for any $\lambda > 0$

$$\Delta_{*,\lambda}^{1-\frac{1}{\alpha}} \left(\log\left(\mathcal{A}T\Delta_*^2\right)\right)^{\frac{1}{2\alpha}} < \lambda^{1-\frac{1}{\alpha}} (\log T)^{\frac{1}{2\alpha}}.$$

Thus, we obtain the following worst-case bound:

$$R_{\mathbb{P}}(T) \leq \frac{K \log(\mathcal{A}T)}{\lambda} + \lambda^{\frac{\alpha-1}{\alpha}} (\log(T))^{\frac{1}{2\alpha}} + \lambda T. \tag{4.27}$$

Now with $\alpha \in [0, \frac{1}{2})$ we consider two different scenarios depending on the relation between K and T. If $K < T^{1-2\alpha}$, then, as one can readily check, by setting in this case $\lambda = T^{-\alpha} \log^{\frac{1}{2}}(T) > \frac{1}{2}\sqrt{\frac{e^{1/2}}{T}}$ (which is an admissible choice according to the Theorem 4.3.8), we obtain:

$$R_{\mathbb{P}}(T) \leq C_1 T^{1-\alpha} (\log T)^{\frac{1}{2\alpha}},$$

where C_1 is some numerical constant.

When $K > T^{1-2\alpha}$ by setting $\lambda = \sqrt{\frac{K}{T}} > \frac{1}{2}\sqrt{\frac{e^{1/2}}{T}}$, we get the following bound:

$$R_{\mathbb{P}}(T) \leq C_2 \sqrt{TK} \log(T),$$

as in this case the second term from Equation (4.27) dominates the bound. Combining these results and taking $C_3 = \max\{C_1, C_2\}$ we obtain the claim of the Theorem. □

4.7.4 Proof of Proposition 4.4.1

Without loss of generality, we suppose that random rewards are bounded in $[-1, 1]$. Recall that we use $\{K\} = \{1, \ldots, K\}$. For $0 \leq i \leq K$, we construct the stochastic bandit environment $\mathcal{B}_i$ in the following way.

For the arm $a \in \{K\}$ we consider the bandit instance $\mathcal{B}_a = \nu_1^a \otimes \ldots \nu_K^a$, where we set

$$\nu_i^a = m_0 \left(\mathcal{R}\left(\frac{1}{2}\right) \mathbb{I}_{a \neq i} + \mathcal{R}\left(\frac{1}{2} + \varepsilon\right) \mathbb{I}_{a=i} \right),$$

where $\mathcal{R}(p)$ is Rademacher distribution with parameter p, i.e. $\mathcal{R}(p) = (1-p)\delta_{-1} + p\delta_1$; $\varepsilon = 1/8$ and m_0 are set to be $T^{-\alpha}$. In every bandit i for each arm $a \in \{K\}$, we assume that the sample rewards are "frozen" from the beginning and drawn from the distribution ν_i^a. More precisely, for the process (X_t^a) attached to the arm a in bandit $\mathcal{B}_i$ we define $X_{t+\ell}^a(\omega) = X_t^a(\omega) \sim \nu_i^a$ for every $\ell \leq T$, $i \in \{0, \ldots, K\}, a \in \{K\}$. One can readily check that process X_t^a satisfies Definition 2.2.2 with rate $\phi_{\mathcal{C}}(t) = 2t^{-\alpha}$. Furthermore, for a sample $X_t^a \sim \nu_i^a$ from arm a in bandit $\mathcal{B}_i$ we have $\mathbb{E}[X_t^a] = 2\varepsilon m_0 = \frac{1}{4}T^{-\alpha}$ if $a = i$, and $\mathbb{E}[X_t^a] = 0$ otherwise.

For the arm a define the following event:

$$E_a = \{N_a \leq cT\},$$

where $c > 0$ is some small universal constant and $N_a = \sum_{t=1}^T \mathbb{I}_{I_t = a}$. We have that under bandit instance $\mathbb{P}_{\mathcal{B}_0}$:

$$T = \mathbb{E}_{\mathcal{B}_0}\left[\sum_{a=1}^K N_a\right] \geq \sum_{a=1}^K \mathbb{E}[N_a | E_a^c] \mathbb{P}_{\mathcal{B}_0}[E_a^c] \geq cTK \min_{a \in \{K\}} \mathbb{P}_{\mathcal{B}_0}[E_a^c].$$

Now since $\mathbb{P}_{\mathcal{B}_0}[E_a^c] = 1 - \mathbb{P}_{\mathcal{B}_0}[E_a]$, we have that

$$\max_{a\in\{K\}} \mathbb{P}_{\mathcal{B}_0}[E_a] \geq 1 - \frac{1}{cK} \geq 1 - \frac{1}{2c},$$

where the last inequality holds since $K \geq 2$.

Denote $a_0 = \operatorname*{Arg\,Max}_{a\in\{K\}} \mathbb{P}_{\mathcal{B}_0}[E_a]$. For the event E_{a_0} in bandit $\mathcal{B}_{a_0}$, by using the change of measure principle between two Rademacher distributions we have:

$$\begin{aligned}
\mathbb{P}_{\mathcal{B}_{a_0}}[E_{a_0}] &= \mathbb{E}_{\mathcal{B}_0}\left[\mathbb{I}_{E_{a_0}} \exp\left(\frac{X_t}{2m_0}\log\left(\frac{1+2\varepsilon}{1-2\varepsilon}\right) + \frac{1}{2}\log\left(\frac{1+2\varepsilon}{1-2\varepsilon}\right)\right)\right] \\
&\geq \mathbb{E}_{\mathcal{B}_0}\left[\mathbb{I}_{E_{a_0}} \exp\left(\frac{-m_0}{2m_0}\log\left(\frac{1+2\varepsilon}{1-2\varepsilon}\right) + \frac{1}{2}\log\left(\frac{1+2\varepsilon}{1-2\varepsilon}\right)\right)\right] \\
&= \mathbb{P}_{\mathcal{B}_0}[E_{a_0}] \geq 1 - \frac{1}{2c}.
\end{aligned}$$

Therefore, for the regret under bandit $\mathcal{B}_{a_0}$ we get

$$\begin{aligned}
\mathbb{E}_{\mathcal{B}_{a_0}}[R_{\mathbb{P}}(T)] &\geq \mathbb{E}_{\mathcal{B}_{a_0}}[R_{\mathbb{P}}(T)|E_{a_0}]\mathbb{P}_{\mathcal{B}_{a_0}}[E_{a_0}] \geq T(1-c)2\varepsilon m_0 \mathbb{P}_{\mathcal{B}_{a_0}}[E_{a_0}] \\
&\geq \frac{1-c}{4}\left(1 - \frac{1}{2c}\right)T^{1-\alpha} \geq \frac{\left(\sqrt{2}-1\right)}{8}T^{1-\alpha},
\end{aligned}$$

which implies the bound on the minimax regret.

Chapter 5

Concentration inequalities for weakly-dependent stationary random fields

In this chapter we consider the interesting phenomenon of concentration of weakly-dependent random fields of possibly high dimension. More precisely we posit a general weak-dependency assumption of a projective kind which can be seen as an extension of the mixingale concept to the case with $d \geq 2$. Based on the martingale-approximation technique and a tree-like ordering of the elements of the integer grid in high dimension, we develop a toolbox (in terms of exponential concentration inequalities and type of Burkholder's inequality). These results extends theoretical results due to Dedecker (1991) for dependent random fields. The results of this chapter is a joint work with Gilles Blanchard and Alexandra Carpentier.

5.1 Introduction

The problem of establishing concentration inequalities for the functionals of stochastic process $(X_t)_{t\in\mathcal{T}}$, distributed according measure $\mathbb{P}$ over its sample space becomes more complicated when the measure $\mathbb{P}$ is not a product measure. In this case one needs to quantify the dependency between the marginals of $\mathbb{P}$. As mentioned in Chapter 2 several works have been done toward the study of the dependent case - either through introducing the notion of mixing coefficients Rosenblatt (1956),Ibragimov (1959),Bradley et al. (1987), or the so-called weak-dependency assumption (Dedecker, 1991; Dedecker et al., 2007). The problem of establishing even the asymptotic results under dependency conditions is already interesting for $Z := \sum_{t\in\mathcal{T}} X_t$, where $\mathcal{T}$ is some subset of some vector space and $(X_t)_{t\in\mathcal{T}}$ is a stochastic process on $(\Omega, \mathcal{F}, \mathbb{P})$, valued in some normed space $(\mathcal{B}, \|\cdot\|)$. When $\mathcal{T} \subset \mathbb{Z}$ we recover the case of partial sums of stochastic processes, while in case $\mathcal{T} \subset \mathbb{Z}^d$ (or more general $\mathcal{T} \subset \mathcal{X}$, where $\mathcal{X}$ is some vector space) we recover the case of $\mathcal{B}-$ valued random fields. We refer to case when $\mathcal{T}$ is some general subspace of a metric space as to the case of $\mathcal{B}-$valued random fields.

In this chapter we study the properties of the deviations of partial sums of real-valued random fields $(X_t)_{t\in\mathcal{T}}$ where $\mathcal{T} \subset \mathbb{N}^d$, $\mathcal{T}$ is a finite rectangle. Many works have been devoted to the study of the asymptotic properties of functions of random fields, and in particular functional central limit theorems, and instance of the law of iterated logarithm, were considered when $\mathcal{T}$ is a rectangle in $\mathbb{N}^d$. Typically, existing works either consider the case where the underlying random field is a (non-linear) transformation of some i.i.d. random field - see Hannan (1973),Wu (2005) - or study the case of a general random field for which a martingale-type assumption on the structure of the underlying $\sigma-$field is used. There are various ways to define martingales in dimension larger than 1 - see e.g. Nahapetian and Petrosian

(1992), Cairoli (1969). A common approach is to consider a stationary random field with a so-called commuting filtration assumption - originally introduced in Fazekas (1983). In this direction, CLT-type results for random fields which are (nonlinear) transformations of i.i.d. random fields are studied in Jirak (2016) for the case $d = 1$ and Giraudo (2018). Furthermore, the weak invariance principle and the law of iterated logarithm are studied in Giraudo (2020). The CLT type results under the assumption that the underlying process consists of pairwise mixing martingale differences - which is satisfied under projective-type of weak-dependency assumption given in this work - is established in Peligrad and Utev (1997). Among many types of conditions the type of projective L_p-condition (Dedecker (1991)) is, to the best of our knowledge, the most general type of assumption when the underlying random field is stationary. Concentration and asymptotic properties of the weighted sums of random fields under this condition are studied in Theorem 1 and Proposition 2 in Dedecker (1991). In that work, a type of martingale-approximation principle is used while using reordering of the random field in $\mathbb{Z}^d$ - which is in some sense unavoidable for martingales in $d > 1$. In the case $d = 1$ a general result (a Burkholder-type inequality) is obtained for stationary stochastic processes in Peligrad et al. (2006) where the bound is expressed in terms of conditional expectation of sums with respect to increasing fields of $\sigma-$algebras.

To the best of our knowledge, in dimension $d > 1$ Burkholder's and Azuma-Hoeffding type inequalities of the paper Dedecker (1991) (Proposition $1a$) and Corollary $3a$) therein) provide the best known upper bounds in the case of general random fields. We notice that sharp Burkholder's bounds and exponential deviation bounds in case $d = 1$ are established in Theorem 1 and Proposition 2 under L_p-projective type assumption for conditional expectations in Peligrad et al. (2007). In our work we consider a stronger projective type assumption than in Dedecker (1991) and under this assumption we prove sharp exponential deviation inequalities and L_p bounds for partial sums of random fields indexed by rectangles in $\mathbb{N}^d$.

To specify the setting, let $\mathcal{R} = \prod_{i=1}^{d}\{0, \dots, n_i\}$ be the $d-$dimensional rectangle with lower left corner in $(0, 0, \dots, 0)$. We pose the definition of projective weak-dependency assumption below.

Assumption 1. Consider a random field $(X_t)_{t\in\mathbb{N}^d}$ $\mathcal{R} = \prod_{i=1}^{d}\{0, \dots, n_i\}$, $p \in [2, +\infty]$, $n_i \in \mathbb{N}$ be defined over probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We assume that it satisfies the following weak-dependency assumption :

$$\|\mathbb{E}[X_t|\mathcal{M}_{t,r}] - \mathbb{E}[X_t]\|_p \leq M_p\varphi_p(r), \tag{5.1}$$

where $\mathcal{M}_{t,r} := \sigma\{X_u : \|u - t\|_{d,\infty} \geq r\}$, $r \geq 0$, $\varphi_p(\cdot)$ is a non-increasing function, $\varphi_p(0) = 1$ and $M_p = \|X_t - \mathbb{E}[X_t]\|_p$. We say that the random field is polynomially weakly-dependent if Equation (5.1) holds with $\varphi_p(r) = r^{-\alpha}$and that it is exponentially weakly-dependent if Equation (5.1) holds with $\varphi_p(r) = \exp(-\gamma r)$, where $\gamma > 0$ is some constant.

Definition 1 defines the dependence coefficient which decays as the distance between X_t and the information 'available' through the conditioning - represented by the $\sigma-$ field $\mathcal{M}_{t,r}$) - increases. In the case of a discrete stochastic process (which can be seen as a the random field in dimension 1) Definition 1 is an extension of a projective type of dependence assumption - see ex. Dedecker and Merlevede (2015) - for $d \geq 2$. Note that the weak-dependency condition 5.1 with $p \geq 2$ implies that the field $(X_t)_{t\in\mathcal{D}}$ is bounded in L_p-norm.

5.1.1 Overview of the main results of the chapter

In this chapter we discuss two main contributions. On the one side we provide non-asymptotic high-probability tail upper bounds for the deviations of partial sums of random fields. On the other side we derive Burkholder-type inequality for the $p$norm, $p \geq 2$ of the partial sums $S_{\mathcal{R}} := \sum_{t\in\mathcal{R}} X_t$. In both cases the bound is expressed as a multi-scale expansion which depends (in terms of a multiplicative constant) on the weak-dependency coefficient based on the distance between consecutive elements on every scale. In the case when $\varphi_p(r) = r^{-\alpha}$ and the random field is defined over the cube $\mathcal{D} = \{n\}_0^d :=$

$\{0,\dots,n-1\}^d$ In the case when $\varphi_p(r) = r^{-\alpha}$ and the random field is defined over the cube $\mathcal{D} = \{n\}^d := \{0,\dots,n-1\}^d$, weak-dependency assumption 5.1 is satisfied for $p=\infty$ we have the following result. For any $\delta \in (0,1]$ with probability at least $1-\delta$ it holds

$$\left| N^{-1} \sum_{t\in\mathcal{D}} (X_t - \mathbb{E}[X_t]) \right| \le C_{\alpha,d} M_\infty N^{-\left(\frac{1}{2}\wedge\frac{\alpha}{d}\right)} K_N \sqrt{\log(1/\delta)}, \tag{5.2}$$

where $N = n^d$, $C_{\alpha,d}$ is some numerical constant and $K_N = 1 + \mathbb{I}_{\alpha=d/2} \log_2^{d/2}(n)(\ln(\log_2(n)))^{d/2}$. If weak-dependency assumption (5.1) holds with some $2 \le p < \infty$, we have

$$\left\| N^{-1} \sum_{t\in\mathcal{D}} (X_t - \mathbb{E}[X_t]) \right\|_p \le C_{\alpha,d} M_p \sqrt{p} K_N N^{-\left(\frac{1}{2}\wedge\frac{\alpha}{d}\right)}. \tag{5.3}$$

This chapter is organized as follows. In Section 5.2 we introduce the setting and necessary notation. Our main results are stated and proved is Section 5.3. In Section 5.4 we compare them to the known bounds under different dependence measures. The multidimensional hierarchical martingale construction and proofs of supplementary technical lemmata are postponed to the last section.

5.2 Notations

Let $\mathcal{D} \subset \mathbb{N}^d$ be a finite subset of $\mathbb{N}^d$ endowed with the standard supremum norm, i.e. for $x, z \in \mathcal{D}$ we write $\|x-y\|_{d,\infty} = \max_{1\le i\le d}|x_i - y_i|$. For a stochastic process $(X_t)_{t\in\mathcal{D}}$ we consider its canonical version over the probability space $\left(\mathbb{R}^{\mathcal{D}}, \mathcal{B}\left(\mathbb{R}^{\mathcal{D}}\right), \mathbb{P}\right)$; in this case X is the identity map $X : \mathbb{R}^{\mathcal{D}} \mapsto \mathbb{R}^{\mathcal{D}}$ and $X_t(\omega)$ is a projection $\mathbb{R}^{\mathcal{D}} \mapsto \mathbb{R}$ on the coordinate $t \in \mathcal{D}$. We refer to the set of all coordinate projections $(X_t)_{t\in\mathcal{D}}$ as to the *random field*. We denote $S_{\mathcal{D}} := \sum_{t\in\mathcal{D}} X_t - \mathbb{E}[X_t]$ for the centered partial sum of $(X_t)_{t\in\mathcal{D}}$. We use w.p. as the short form of "with probability" and $a\wedge b$, $a\vee b$ to denote minimum and maximum of any real numbers a, b correspondingly. As usual under $\lceil a\rceil$, $[a]$,$\lfloor a\rfloor$ we understand ceil, integer and floor part of a number $a \in \mathbb{Z}$. For any $k \in \mathbb{N}$ we denote $\{k\}_0 := \{0,\dots,k-1\}$ if $k \ge 1$ and $\{k\}_0 = \{\emptyset\}$ if $k = 0$; denote also $\{k\} := \{1,\dots,k\}$ for $k \ge 1$. Furthermore, for a $d-$dimensional set $A \subset \mathbb{N}^d$, we denote its cardinality as $|A|$. For a set $A \subset \mathbb{N}^d$, $a \in \mathbb{R}$, $b \in \mathbb{N}^d$ under aA we understand the set $\{ae : e \in A\} \subset \mathbb{N}^d$ and under $b + A$ we understand the set $\{b + e : e \in A\} \subset \mathbb{N}^d$. Finally, for any collection of sets $(A_t)_{t\in\mathbb{T}}$, $A_t \subset \mathbb{N}^d$ for any $(i,j) \in \mathbb{T}^{\otimes 2}$ under $A_i + A_j$ we understand the set $\{e_1 + e_2 : e_1 \in A_i, e_2 \in A_j\}$ and if the collection $(A_t)_{t\in\mathbb{T}}$ is disjoint, so under $\biguplus_{t\in\mathbb{T}} A_t$ we understand their disjoint union.

5.3 Main results

In this section we prove the main results for the case of random fields indexed by an arbitrary rectangle in $\mathbb{N}^d$.

Theorem 5.3.1. *Let $\mathcal{R} = \prod_{i=1}^d \{N_i\}_0$ be a d-dimensional rectangle of sidelengths $N_i \ge 1$, $i \in \{1,\dots,d\}$, and $m(\mathcal{R}) := \max_{i=1,\dots,d}\lfloor \log_2 N_i \rfloor$. Let $\boldsymbol{\delta} = (\delta_k)_{k\ge 1}$ be a fixed sequence of integers such that*

$$\sum_{k=1}^{m(\mathcal{R})} \delta_k 2^{-k} \le \frac{1}{4d^2}. \tag{5.4}$$

Let $(X_t)_{t\in\mathbb{N}^d}$ be a random field such that Assumption 1 is satisfied for some $p \in [2,\infty]$.

- *if $p \in [2,\infty)$, then*

$$\|S_{\mathcal{R}}\|_p \le C_{p,d} \Psi_p(\boldsymbol{\delta}, \mathcal{R}),$$

with $C_p = 4\sqrt{p}$*;*

- *if* $p = \infty$*, then for any* $\delta \in (0, 1]$ *with probability at least* $1 - \delta$*:*

$$|S_{\mathcal{R}}| \leq C_\infty \Psi_p(\boldsymbol{\delta}, \mathcal{R}) \sqrt{\log\left(\frac{1}{\delta}\right)},$$

with $C_\infty := 10$*,*

where (putting $\delta_0 = 0$*)*

$$\Psi_p(\boldsymbol{\delta}, \mathcal{R}) := 2M_p \sqrt{|\mathcal{R}|}\big(1 + \varphi_p(1) + \sum_{k=1}^{m(\mathcal{R})+1} \varphi_p\big(\delta_{k-1} + 1\big) \sqrt{|\mathcal{C}_{k,0} \cap \mathcal{R}|}\big),$$

with $\mathcal{C}_{k,0} := \big\{2^k\big\}_0^d$.

Proof We proof the claim by induction on the size of the rectangle $\mathcal{R}$. We consider the case $p < \infty$ only; the arguments for the case $p = \infty$ are the same and follow by exchanging the $p-$norm by subgaussian norm (i.e. the quantity which is defined as $\|X\|_{SG} := \inf_{c>0}\{\mathbb{E}[\exp(\lambda X)] \leq \exp\Big(\frac{\lambda^2 c^2}{2}\Big), \lambda > 0\}$) and using convertion from subgaussian norm to the exponential probability bound. For $\mathcal{R}$ reduced to a single element $\mathbf{0}$, the claim obviously holds. Now, assume the claim is established for any rectangle $\mathcal{R}' \subsetneq \mathcal{R}$. We use the following construction to obtain the sharp bound for the deviations of partial sums of the processes. For any integer $\delta < 2^k$ let

$$\Lambda_{k,\delta} := 2^k \mathbb{N}_{>0} + \{\delta\}_0. \tag{5.5}$$

Let $\boldsymbol{\delta} = (\delta_k)_{k \geq 1}$ be a fixed sequence of integers with $\delta_k \leq 2^k$, $k \geq 1$. Define

$$\Lambda_{\boldsymbol{\delta}} := \bigcup_{k \geq 1} \Lambda_{k,\delta_k} \qquad \mathfrak{F}_{\boldsymbol{\delta}} := (\mathbb{N} \setminus \Lambda_{\boldsymbol{\delta}})^d; \qquad \mathfrak{F}_{\boldsymbol{\delta}}^c := \mathbb{N}^d \setminus \mathfrak{F}_{\boldsymbol{\delta}}. \tag{5.6}$$

We call $\mathfrak{F}_{\boldsymbol{\delta}}$ the "framed set" and $\mathfrak{F}_{\boldsymbol{\delta}}^c$ the "frame". For the set $A \subset \mathbb{N}^d$ we use the decomposition into $\mathcal{R} = \mathcal{R} \cap (\mathfrak{F}_{\boldsymbol{\delta}} \cup \mathfrak{F}_{\boldsymbol{\delta}}^c)$ and by triangle inequality

$$\|S_{\mathcal{R}}\|_p \leq \|S_{\mathcal{R} \cap \mathfrak{F}_{\boldsymbol{\delta}}}\|_p + \left\|S_{\mathcal{R} \cap \mathfrak{F}_{\boldsymbol{\delta}}^c}\right\|_p.$$

By Proposition 5.5.5 we have $\|S_{\mathcal{R} \cap \mathfrak{F}_{\boldsymbol{\delta}}}\|_p \leq \frac{C_{p,d}}{2} \Psi_p(\boldsymbol{\delta}, \mathcal{R})$. For the second term, we first decompose $\mathfrak{F}_{\boldsymbol{\delta}}^c$ as a disjoint union of product sets, writing $\Lambda_{\boldsymbol{\delta}}^c := \mathbb{N} \setminus \Lambda_{\boldsymbol{\delta}}$, as:

$$\mathfrak{F}_{\boldsymbol{\delta}}^c = \mathbb{N}^d \setminus (\Lambda_{\boldsymbol{\delta}}^c)^d = \biguplus_{i=1}^d \Delta_i, \qquad \Delta_i := \left(\prod_{j=1}^{i-1} \Lambda_{\boldsymbol{\delta}}^c \times \Lambda_{\boldsymbol{\delta}} \times \prod_{j=i+1}^{d} \mathbb{N}\right),$$

therefore, by the triangle inequality,

$$\left\|S_{\mathcal{R} \cap \mathfrak{F}_{\boldsymbol{\delta}}^c}\right\|_p = \left\|\sum_{i=1}^d S_{\mathcal{R} \cap \Delta_i}\right\|_p \leq \sum_{i=1}^d \|S_{\mathcal{R} \cap \Delta_i}\|_p. \tag{5.7}$$

We introduce the following notation. For a set $B \subset \mathbb{N}$, and an integer $j \in \{|B|\}_0$, denote $j : B$ the $(j+1)$-th element of B in increasing order. For a product set $\mathbf{A} = \prod_{i=1}^d A_i$, and a d-tuple $\boldsymbol{t} =$

$(t_1, \dots, t_d) \in \mathcal{K}(\mathbf{A}) := \prod_{i=1}^d \{|A_i|\}_0$, denote $t : \mathbf{A} = (t_1 : A_1, \dots, t_d : A_d)$, and the "compressed" version of the restriction of the process $(X_t)_{t\in\mathbb{N}^d}$ to $\mathcal{K}(\mathbf{A})$ as

$$\widetilde{X}_t^{(\mathbf{A})} = X_{t:\mathbf{A}}, \qquad t \in \mathcal{K}(\mathbf{A}) \qquad \widetilde{S}_{\mathcal{K}(\mathbf{A})}^{(\mathbf{A})} := \sum_{t\in\mathcal{K}(\mathbf{A})} \widetilde{X}_t^{(\mathbf{A})} = S_{\mathbf{A}}. \tag{5.8}$$

Since Δ_i is a product set, so is $\mathcal{R} \cap \Delta_i$, and we can apply the above '"compression principle"'. Using assumption (5.4), the side-length of the compressed version of $\mathcal{R} \cap \Delta_i$ along direction i is bounded by

$$|\Lambda_{\boldsymbol{\delta}} \cap \{N_i\}_0| \le \left| \bigcup_{k\ge 1} \Lambda_{k,\delta_k} \cap \{N_i\}_0 \right| \le \sum_{k=1}^{\lfloor \log_2(N_i)\rfloor} \left\lfloor \frac{N_i}{2^k} \right\rfloor \delta_k \le N_i \sum_{k=1}^{m(\mathcal{R})} 2^{-k}\delta_k \le \frac{N_i}{4d^2},$$

while for $j \neq i$ the side-lengths are bounded by N_j, hence

$$|\mathcal{R} \cap \Delta_i| \le \frac{|\mathcal{R}|}{4d^2}. \tag{5.9}$$

By Lemma 5.5.6 process $\left(\widetilde{X}_t^{(\mathcal{R}\cap\Delta_i)}\right)_{t\in\mathbb{N}^d}$ satisfies weak-dependency condition (5.1). Applying the induction hypothesis to the process $\left(\widetilde{X}_t^{(\mathcal{R}\cap\Delta_i)}\right)_{t\in\mathbb{N}^d}$ over the rectangle $\mathcal{K}(\mathcal{R} \cap \Delta_i)$, we obtain

$$\|S_{R\cap\Delta_i}\|_p = \left\|\widetilde{S}_{R\cap\Delta_i}^{(R\cap\Delta_i)}\right\|_p \le C_p \Psi_p(\boldsymbol{\delta}, \mathcal{K}(\mathcal{R} \cap \Delta_i)). \tag{5.10}$$

We estimate this upper bound using (5.9) and straightforward cardinality bounds via:

$$\begin{aligned}
\Psi_p(\boldsymbol{\delta}, \mathcal{K}(\mathcal{R} \cap \Delta_i)) &= 2M_p\sqrt{|\mathcal{R} \cap \Delta_i|}\left(1 + \varphi_p(1) + \sum_{k=1}^{m(\mathcal{K}(\mathcal{R}\cap\Delta_i))+1} \varphi_p\big(\delta_{k-1}+1\big)\sqrt{|\mathcal{C}_{k,0} \cap \mathcal{K}(\mathcal{R} \cap \Delta_i)|}\right) \\
&\le \frac{1}{d} M_p\sqrt{|\mathcal{R}|}\left(1 + \varphi(1) + \sum_{k=1}^{m(\mathcal{R})+1} \varphi_p\big(\delta_{k-1}+1\big)\sqrt{|\mathcal{C}_{k,0} \cap \mathcal{R}|}\right) \\
&= \frac{1}{2d}\Psi_p(\boldsymbol{\delta}, \mathcal{R}).
\end{aligned} \tag{5.11}$$

Finally using $\|S_{\mathcal{R}\cap\mathfrak{F}_{\boldsymbol{\delta}}}\|_p \le \frac{C_{p,d}}{2}\Psi_p(\boldsymbol{\delta}, \mathcal{R})$, Proposition 5.5.5, (5.7), (5.10), (5.11), we obtain

$$\|S_{\mathcal{R}}\|_p \le \frac{1}{2}C_p\Psi(\boldsymbol{\delta}, \mathcal{R}) + \sum_{i=1}^d \frac{1}{2d} C_p\Psi_p(\boldsymbol{\delta}, \mathcal{R}) \le C_p\Psi_p(\boldsymbol{\delta}, \mathcal{R}),$$

and the induction claim is proved.

□

Now, for a cube $\mathcal{D} = \{n\}_0^d$, given the particular weak-dependency rate $\varphi_p(t)$ we choose the sequence $\boldsymbol{\delta} = (\delta_k)_{k\ge1}$ such that criterion is fulfilled and the value on the rhs of function $\Psi_p(\delta, \mathcal{R})$ is close to its minimum. This leads to the following result which gives the deviation rates for uniform cubes.

Corollary 5.3.2. *Let $\mathcal{D} = \{n\}_0^d$ be a $d-$dimensional cube with the side-length n and $m(\mathcal{D}) = \lfloor \log_2 n \rfloor$. Consider a random field $(X_t)_{t\in\mathbb{N}^d}$ which satisfies the weak-dependency Assumption 1 for a given $p \in [2, +\infty]$ with the rate $\varphi_p(\cdot)$. Then*

- *If $p \in [2, +\infty)$ and rate $\varphi_p(t) = t^{-\alpha}$, $\alpha > 0$,*

$$\|S_{\mathcal{D}}\|_p \le C_{p,d}^{(1)} M_p n^{d-\left(\frac{d}{2}\wedge\alpha\right)} \text{ if } \alpha \neq d/2 \qquad \|S_{\mathcal{D}}\|_p \le C_{p,d}^{(2)} M_p n^{\frac{d}{2}} (\log_2 n)^{\frac{d}{2}} \text{ if } \alpha = d/2, \tag{5.12}$$

where $C_{p,d}^{(1)} := 4\sqrt{p}\Big(1+\varphi(1)+2^{\frac{d}{2}}+\frac{d^{2\alpha}2^d}{\big(1-2^{\frac{d/2-\alpha}{1+\alpha}}\big)^{1+\alpha}}\Big)$, $C_{p,d}^{(2)} := 4\sqrt{p}\Big(2^{d/2}+\big(2\sqrt{2}\big)^d\Big)$

- *If* $p=\infty$ *then we have*

$$\|S_{\mathcal{D}}\|_{SG} \le C_{p,d}^{(1)} M_p n^{d-\left(\frac{d}{2}\wedge\alpha\right)} \text{ if } \alpha \neq d/2 \qquad \|S_{\mathcal{D}}\|_{SG} \le C_{p,d}^{(2)} M_p n^{\frac{d}{2}} (\log_2 n)^{\frac{d}{2}} \text{ if } \alpha = d/2, \tag{5.13}$$

Furthermore, if the decay rate of weakly-dependent coefficients is $\phi_p(t)=\exp(-\gamma t^\eta)$ *then we have the rate* $\|S_{\mathcal{D}}\|_p \le M_p C_{p,d,\gamma,\eta} n^{d/2}$*, where* $p\in[2,+\infty]$ *and some numerical constant* $C_{p,d,\gamma,\eta}>0$*.*

Proof Notice that for the cube $\mathcal{D}=\{n\}_0^d$ as a direct consequence from Theorem 5.3.1 we have that for the rate $\varphi_p(t)=t^\alpha$ it is sufficient to choose $\boldsymbol{\delta}$ such that constraints 5.4 is fulfilled and $\Psi_p(\boldsymbol{\delta},\mathcal{D}) := M_p\sqrt{n^d}\Big(1+\sum_{k=1}^{\lfloor\log_2 n\rfloor+1}(\delta_{k-1}+1)^{-\alpha}\sqrt{2^{kd}}\Big)$, $\delta_0=1$ is close to its minimum. Using Lagrange multiplier method to solve the constrained optimization problem

$$\min_{\left(\widetilde{\delta}_k\ge 1\right)} \sum_{k=1}^m \widetilde{\delta}_k^{-\alpha} 2^{\frac{kd}{2}}, \text{ s.t. } \sum_{k=1}^m \frac{\widetilde{\delta}_k}{2^k} \le \frac{1}{4d^2}$$

where $m=\lfloor\log_2 n\rfloor$, we obtain $\widetilde{\delta}_j = \frac{2^{j\frac{\frac{d}{2}+1}{1+\alpha}}}{4d^2\sum_{k=1}^m 2^{\frac{k\left(\frac{d}{2}-\alpha\right)}{1+\alpha}}}$, $1\le j\le m$. Taking $\delta_j := \lfloor\widetilde{\delta}_j\rfloor$ one readily checks that Equation (5.4) is satisfied. Furthermore, with this choice of $\boldsymbol{\delta}=(\delta_j)_{j\ge1}$ we have

$$\Psi_p(\boldsymbol{\delta},\mathcal{D}) \le 2M_p\sqrt{n^d}\Big(1+2^{\frac{d}{2}}\Big(1+\big(4d^2\big)^\alpha\Big(\sum_{k=1}^m 2^{\frac{k\left(\frac{d}{2}-\alpha\right)}{1+\alpha}}\Big)^{1+\alpha}\Big)\Big)$$

Now if $\alpha>\frac{d}{2}$ then $\Psi_p(\boldsymbol{\delta},\mathcal{D}) \le 2M_p\sqrt{n^d}\Big(1+\varphi(1)+2^{\frac{d}{2}}\Big(1+\frac{d^{2\alpha}2^{\frac{d}{2}}}{\big(1-2^{\frac{d/2-\alpha}{1+\alpha}}\big)^{1+\alpha}}\Big)\Big)$.

If $\alpha<\frac{d}{2}$ then

$$\Psi_p(\boldsymbol{\delta},\mathcal{D}) \le 2M_p\sqrt{n^d}\Big(1+2^{\frac{d}{2}}\Big(1+\varphi(1)+\frac{d^{2\alpha}2^{\frac{d}{2}+\alpha}}{\big(2^{\frac{d/2-\alpha}{1+\alpha}}-1\big)^{1+\alpha}}n^{\frac{d}{2}-\alpha}\Big)\Big) \le M_p n^{d-\alpha}C_d,$$

where $C_d := \Big(1+2^{\frac{d}{2}}(1+\varphi(1))+\frac{d^{2\alpha}2^{d+\alpha}}{\big(2^{\frac{d/2-\alpha}{1+\alpha}}-1\big)^{1+\alpha}}\Big)$.

Lastly, in the case $\alpha=\frac{d}{2}$ we get

$$\Psi_p(\boldsymbol{\delta},\mathcal{D}) \le 2M_p 2^{\frac{d}{2}}\big(1+\varphi(1)+(2d)^d\big)n^{\frac{d}{2}}(\log_2(n))^{\frac{d}{2}+1}$$

Finally the calculations for the case $p=\infty$ are identical and the claim follows. □

5.4 Discussion

In this section we compare bounds from Proposition 5.3.1 to the analogous results for the non i.i.d. real random fields. In many of such results (see ex. (Dedecker, 1991; Doukhan et al., 1984; Rio, 2000)) exponential inequality for partial sums of bounded random fields is derived from $L_p(\mathbb{P})$ bound by optimizing over the value $p\ge 2$. In general, the analysis of the properties of the random fields are mostly based

either on the notion of multidimensional martingale or on the coupling with the (nonlinear) transform of the i.i.d. random field. In our examples below we discuss methods based on the multidimensional martingale and provide several references to the works on coupling argument.

Notice that weak-dependency condition (5.1) in the case $d = 1$ is stronger as the so-called "mixingale type" condition (see for example Mc Leish (1975)). The former is mentioned in the works Dedecker et al. (2007), Dehling and Philipp (1982) to characterise fading correlation between the past and the future of a discrete stochastic process.

Example 5.4.1. Multidimensional martingales and switching filtrations

The notion of the multidimensional martingale is based on the so-called commuting filtration; it was introduced in Fazekas (1983). We recall it below by introducing the following notation.

For $k, \ell \in \mathbb{Z}^d$ we say that $k \leq_{cw} \ell$ if $k_i \leq \ell_i$ for every $i \leq d$. As before we assume we work over some probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and say that the sequence of sigma-fields $(\mathcal{F}_\ell)_{\ell \in \mathbb{Z}^d} \subset \mathcal{A}$ is a filtration if $\mathcal{F}_k \subset \mathcal{F}_\ell$ for $k \preceq \ell$. We say that the filtration $(\mathcal{F}_\ell)_{\ell \in \mathbb{Z}^d}$ is *commuting* if for every $Y \in L_1(\mathbb{P})$ it holds $\mathbb{E}[\mathbb{E}[Y|\mathcal{F}_k]|\mathcal{F}_\ell] = \mathbb{E}[\mathbb{E}[Y|\mathcal{F}_\ell]|\mathcal{F}_k] = \mathbb{E}\left[\mathbb{E}\left[Y|\mathcal{F}_{\min(k,\ell)}\right]\right]$, where $\min(k, \ell) := (\min(k_i, \ell_i))_{i \leq d}$. Denote $\mathbf{n} = (n, \dots, n)$. A collection of random variables $(Z_\ell)_{\ell \in \mathbb{Z}^d}$ is called an *orthomartingale* random field with respect to a commuting filtration $(\mathcal{F}_\ell)_{\ell \in \mathbb{Z}^d}$ if Z_n is $\mathcal{F}_n-$measurable for every $n \in \mathbb{Z}^d$ and for any $k, \ell \in \mathbb{Z}^d$ such that $k \preceq \ell$ we have $\mathbb{E}[Z_\ell|\mathcal{F}_k] = Z_k$. Lastly, one also defines the *orthomartingale difference field* $(Y_\ell)_{\ell \in \mathbb{Z}^d}$ if for $k \preceq \ell$ holds $\mathbb{E}[Y_\ell|\mathcal{F}_k] = 0$.

For orthomartingale difference fields the results of Giraudo (2019) and Fazekas (2005) ensure (tight) control of the $p-$ norm of partial sums of random fields $(X_t)_{t \in \mathcal{D}}$ by means of Burkholder's type inequality. Namely for the cube $\mathcal{D} = \{n\}^d$ and orthomartingale difference random field $(Z_k, \mathcal{F}_k)_{k \preceq \mathbf{n}}$ from Theorem 3.1 in Fazekas (2005) one obtains the following upper bound for $p \geq 2$ and martingale $S_\mathbf{n} = \sum_{k \preceq \mathbf{n}} Z_k$

$$\|S_\mathbf{n}\|_p \leq \sqrt{p} n^{\frac{d}{2}} \max_{i \preceq \mathbf{n}} \|Z_i\|_p.$$

Notice that a centered orthomartingale-difference random field $(Z_k, \mathcal{F}_k)_{k \in \mathcal{D}}$ satisfies weak-dependence condition (5.1) as for any $t \in \mathcal{D}$, $r \in \mathbb{N}$ it holds $\mathbb{E}\left[Z_t|\overline{\mathcal{F}}_{B_{t,r}}\right] = 0$ which implies weak-dependence with any rate. Also the constant $\sqrt{p}$ is optimal (see Theorem 4.3 in Pinelis (1994)) and thus it can be used to derive the exponential bounds (by using techniques described in Rio (2000), see also Corollary 3.1 in Dedecker (1991)) of optimal order from the $p-$norm inequality. Furthermore, for the orthomartingale difference random fields with values in an arbitrary Banach space Theorem 1.13 in Giraudo (2019) provides a deviations type result for norms of random sums. For $(X_t, \mathcal{F}_t)_{t \in \mathbb{Z}^d}$ and $S_\mathbf{n} = \sum_{\mathbf{1} \preceq i \preceq \mathbf{n}} X_i$, $p \geq 1$ and $x > 0$ we have that

$$\mathbb{P}\left[\|S_\mathbf{n}\| > x n^{d/p}\right] \leq C \int_0^{+\infty} \mathbb{P}[\|X_1\| > xu] u^{p-1} (1 + |\log(u)|)^{d+1} du.$$

A Burkholder's inequality is then deduced by multipliying both sides with px^{p-1} and integrating with respect to x from 0 to ∞.

Example 5.4.2. Bernoulli random fields

Similarly, an extension of a Burkholder's type of inequality to the processes which can be viewed as a real-valued nonlinear transform of i.i.d. random field is given in El Machkouri et al. (2013). In this work authors consider the so-called Bernoulli random fields of the form $g\left(\varepsilon_{k-s}, s \in \mathbb{Z}^d\right)$, $k \in \mathbb{Z}^d$ where $(\varepsilon_i)_{i \in \mathbb{Z}^d}$ are i.i.d. random variables and g is some measurable function. This class of random fields has been studied recently focusing mainly on the functional central limit theorem framework (see ex. Biermé and Durieu (2014), Klicnarová et al. (2016)) and large deviation principle (see Sang and Xiao (2018)). In Giraudo (2020) authors prove a variant of law of Iterated logarithm. Typical examples which belongs to this class are linear random fields and Volterra random fields (see examples in Sang

and Xiao (2018) and section 2 in Giraudo (2019)). In this case dependence measure condition is the so-called physical dependence measure (originally introduced in Wu (2005)). It is defined for the process $(X_t)_{t\in\mathcal{D}}$, $\mathcal{D}=\{n\}_0^d$, X_t such that $\|X_t\|_p<\infty$ through coupling coefficient $\delta_{i,p}=\|X_i-X_i^*\|_p$, where the coupled version $X_i^*=g(\varepsilon^*_{i-s})$ and $\varepsilon_j^*=\varepsilon_j\mathbb{I}_{j\neq 0}+\varepsilon_0'\mathbb{I}_{j=0}$, with $(\varepsilon_i^*)_{i\in\mathbb{Z}^d}$ being an i.i.d. copy of random field $(\varepsilon_i)_{i\in\mathbb{Z}^d}$. In this framework Proposition 1 in El Machkouri et al. (2013) implies following Burkholder's type inequality

$$\|S_{\mathcal{D}}\|_p=\|S_{\mathbf{n}}\|_p\leq\sqrt{2p}n^{d/2}\Delta_p,$$

where $\Delta_p=\sum_{i\in\mathbb{Z}^d}\delta_{i,p}<\infty$. Although it is strictly speaking incorrect to compare two notions of dependence, in particular case of linear i.i.d. fields this condition is stronger then weak-dependency assumption 5.1.

Example 5.4.3. L_p- **projective criterion**

In the work Dedecker (1991) a Burkholder's type inequality is obtained under the L_p-projective dependence criterion. More precisely, for a centered and square-integrable random field $(X_t)_{t\in\mathcal{D}}$ and $S_{\mathcal{D}}=\sum_{t\in\mathcal{D}}X_t$ one has the following $p-$norm type of inequality

$$\|S_{\mathcal{D}}\|_p\leq\sqrt{2p\sum_{t\in\mathcal{D}}b_{t,p/2}(\mathcal{D})},\tag{5.14}$$

where $b_{t,\beta}=\left\|X_t^2\right\|_\beta+\sum_{k\in V_t^1}\left\|X_k\mathbb{E}_{|k-t|}[X_t]\right\|_\beta$, is finite, $\beta>0$ and V_t^1 denotes the set of all elements which preceed t in lexicographic order and $\mathbb{E}_\ell[X_t]-$ expectation conditioned with respect to $\sigma-$algebra $\mathcal{F}^\ell_{V_t^1}$ generated by all the elements from V_t^1 at distance at least ℓ (in terms of ℓ_∞ norm) from t. Notice that firstly with this definition for every $\ell\in\mathbb{N}$ filtration $\mathcal{F}^\ell_{V_t^1}$ as filtration in t is not commuting so that the correspondent results for orthomartingales cannot be applied. Secondly, condition on the construction of $\sigma-$ algebra $\mathcal{F}^r_{V_i^t}$ is weaker then $\overline{\mathcal{F}}_{B(t,r)}$, thus weak-dependency coefficient $\varphi(r)$ is larger than $\mathbb{E}_{|k-t|}[X_t]$. It is easy to see that if $(X)_{t\in\mathcal{D}}$ is martingale-difference random fields (w.r.t lexicographic ordering) then (5.14) implies a standard Burkholder's inequality for martingales as in this case $b_{t,p/2}=\left\|X_t^2\right\|_{p/2}=\|X_t\|_p^2$. Furthermore, as soon as weak-dependency condition (5.1) holds and $p\geq 2$ we have $\mathbb{E}_{|k-t|}[X_t]_p\leq M_p\varphi_p(|k|)$, so that by using Hölders inequality and regrouping same weak-dependent coefficients $\varphi_p(\cdot)$ we obtain

$$b_{t,p/2}\leq M_p^2\big(1+\sum_{k\in V_t^1}\varphi_p(|k|)\big)=M_p^2\big(1+\sum_{k=1}^{\infty}k^{d-1}\varphi_p(k)\big).$$

The latter inequality ensures the same upper bound as in the case of martingale differences when $\varphi_p(k)\leq k^{-\alpha}$ and $\alpha>d$. This result is worse in the sense that it provides the optimal bound when $\alpha>d$, whereas the bound of Theorem 5.3.2 for cubes is optimal (i.e. of the same order as in the case of martingale differences fields) when $\alpha>\frac{d}{2}$.

5.5 Proofs of the auxiliary results of Chapter 5

The key element of the proof of Theorem 5.3.1 is the tree-like recursive ordering over $\mathbb{N}^d$. To define it we introduce the following notation. For $t=(t_1,\dots,t_d)\in\mathbb{N}^d$ define $\pi_k(t):=\big(\lfloor 2^{-k}t_i\rfloor 2^k\big)_{1\leq i\leq d}\in 2^k\mathbb{N}^d$. Observe that $\pi_0(t)=t$, and that $\pi_k(t)=\underline{0}$ for $k\geq\log_2\|t\|_\infty$. Let $\leq_{\text{lex}}$ denote the lexicographical order on $\mathbb{N}^d$. Denote $<_{\text{lex}}$ to be the associated strict order relation. For two elements t,t' of $\mathbb{N}^d$, define

$$\kappa(t,t')=\min\{k\in\mathbb{N}:\pi_k(t)=\pi_k(t')\}-1.\tag{5.15}$$

Note that $\kappa(t,t')$ is always well-defined, since $\pi_k(t) = \pi_k(t') = \underline{0}$ for $k \geq \max(\|t\|_\infty, \|t'\|_\infty)$, hence the minimum in Equation (5.15) is over a non-empty set. Furthermore $\kappa(t,t') = -1$ iff $t = t'$. We define the following total order $\preceq$ on $\mathbb{N}^d$

$$t \preceq t' \quad \text{iff} \quad \text{either } \kappa(t,t') = -1 \text{ or } \pi_{\kappa(t,t')}(t) \leq_{\text{lex}} \pi_{\kappa(t,t')}(t'). \tag{5.16}$$

It is straightforward to check that $\preceq$ is a total order over $\mathbb{N}^d$ since $\leq_{\text{lex}}$ is a total order over $\mathbb{N}^d$. This order can be described as the co-lexicographical order for the (one-to-one) sequence representation $(\pi_k(t))_{k\geq 0}$ of $t \in \mathbb{N}^d$, where the base order for the elements of the sequence is the usual lexicographical order. Equivalently, this is the co-lexicographical order on the (infinite) binary representation $(\lfloor t_i 2^{-k} \rfloor \bmod 2)_{i=d,\ldots,1})_{k\geq 0}$, where the vectorization is along (reverse) dimension first, then along scale.

For $t \in \mathbb{N}^d$, let

$$\Pi_k^{\prec}(t) := \{t' \in \mathbb{N}^d : \pi_k(t') \prec \pi_k(t)\}, \tag{5.17}$$

$$\Pi_k(t) := \{t' \in \mathbb{N}^d : \pi_k(t') \preceq \pi_k(t)\}, \tag{5.18}$$

where $t \prec t'$ indicates that t is strictly less than t' for the order $\preceq$. Next we need the following Lemma which describes properties of the order $\preceq$.

Lemma 5.5.1. *Let ordering $\leq_{cw}$ be the partial order on $\mathbb{N}^d$ such that for $t = (t_1, \ldots, t_d) \in \mathbb{N}^d$, $t' = (t_1', \ldots, t_d') \in \mathbb{N}^d$ we say that $t \leq_{cw} t'$ iff $t_i \leq t_i'$ for all $i \in \{d\}$. The following statements hold true:*

0) *For any k, ℓ such that $k \leq \ell$ it holds $\pi_\ell \circ \pi_k = \pi_k \circ \pi_\ell = \pi_\ell$.*

i) The partial order $\leq_{\text{cw}}$ is compatible with both the total orders $\leq_{\text{lex}}$ and $\preceq$, meaning that

$$t \leq_{\text{cw}} t' \implies t \leq_{\text{lex}} t' \text{ and } t \preceq t'.$$

ii) For any $t \in \mathbb{N}^d$ and $k, \ell \in \mathbb{N}$ with $k \geq \ell$ it holds $\pi_k(t) \preceq \pi_\ell(t)$. In particular, in case $\ell = 0$ it holds $\pi_k(t) \preceq t$.

iii) All applications π_k are monotone nondrecreasing with respect to $\preceq$:

$$\forall k \in \mathbb{N}, \ \ \forall t, t' \in \mathbb{N}^d : \qquad t \preceq t' \Rightarrow \pi_k(t) \preceq \pi_k(t').$$

iv) For a positive integer k put $\mathcal{C}_{k,0} := \{2^k\}_0^d$, and, for $b \in 2^k\mathbb{N}^d$, put $\mathcal{C}_{k,b} := b + \mathcal{C}_{k,0}$. For any $t \in \mathcal{C}_{k,b}$, it holds $\pi_k(t) = \pi_k(b) = b$, $\Pi_k^{\prec}(t) = \Pi_k^{\prec}(b)$ and $\Pi_k(t) = \Pi_k(b)$.

v) For any $t \in \mathbb{N}^d$ and $k \in \mathbb{N}$, it holds

$$\Pi_k^{\prec}(t) = \{t' \in \mathbb{N}^d : t' \prec \pi_k(t)\}, \tag{5.19}$$

$$\Pi_k(t) = \Pi_k^{\prec}(t) \cup \mathcal{C}_{k,\pi_k(t)}. \tag{5.20}$$

vi) For any $k \in \mathbb{N}_{>0}$ and $t \in \mathbb{N}^d$, it holds

$$\Pi_k^{\prec}(t) \subseteq \Pi_{k-1}^{\prec}(t) \subseteq \Pi_{k-1}(t) \subseteq \Pi_k(t). \tag{5.21}$$

Proof 0. The equality $\pi_k \circ \pi_\ell = \pi_\ell$ follows directly from the definition of the floor part and since $\ell \geq k$. For the second claim notice that from the definition of floor part for any $t \in \mathbb{N}$ we have $\lfloor \frac{\lfloor \frac{t}{2^k} \rfloor 2^k}{2^\ell} \rfloor \leq \lfloor \frac{t}{2^\ell} \rfloor$ which implies $\pi_\ell \circ \pi_k(t) \leq \pi_\ell(t)$. The converse inequality follows from simple inequality $\lfloor ax \rfloor \geq a \lfloor x \rfloor$

which is true for $a \in \mathbb{N}$, $x > 0$ applied with $x = \frac{t}{2^\ell}$ and $a = 2^{\ell-k}$.
i) The implication for the lexicographical order is obvious; concerning the order $\preceq$, note that obviously all the mappings π_k for $k \geq 0$ are non-decreasing for the partial order $\leq_{\rm cw}$, i.e. $t \leq_{\rm cw} t'$ implies $\pi_k(t) \leq_{\rm cw} \pi_k(t')$, in turn implying $\pi_k(t) \leq_{\rm lex} \pi_k(t')$ for all k, which finally entails $t \preceq t'$ from the definition.

ii): The second claim follows directly from **i**) since it follows from the definition that $\pi_k(t) \leq_{\rm cw} \pi_\ell(t)$ if $k \geq \ell$.

iii) Assume $t \prec t'$ and let $\kappa = \kappa(t, t') \geq 0$. Then by definition of $\pi_\ell(\cdot)$, for all $\ell > \kappa$ it holds $\pi_\ell(t) = \pi_\ell(t')$, while $\pi_\kappa(t) <_{\rm lex} \pi_\kappa(t')$. Thus, for $k > \kappa$ it holds $\pi_k(t) \preceq \pi_k(t')$; for $k = \kappa$ it holds $\pi_k(t) \preceq \pi_k(t')$ from **i**); and for $k < \kappa$, for any $\ell \geq \kappa$ we have $\pi_\ell \circ \pi_k = \pi_\ell$, so the conditions for $\pi_k(t) \prec \pi_k(t')$ are met. In both cases we have $\pi_k(t) \preceq \pi_k(t')$.

iv) For $u \in 2^k\mathbb{N}$, it holds $\lfloor 2^{-k}(u+v) \rfloor = u$ iff $v \in \{2^k\}_0$. It follows that for $t, t' \in \mathbb{N}^d$, $\pi_k(t) = \pi_k(t')$ iff $t' \in \mathcal{C}_{k,\pi_k(t)}$. The claims follow from the definitions of $\pi_k, \Pi_k^\prec$ and Π_k.
v) For $t, t' \in \mathbb{N}^d$, if $t' \prec \pi_k(t)$ then $\pi_k(t') \preceq t' \prec \pi_k(t)$, from **ii**). Conversely, if $t' \succeq \pi_k(t)$, then $\pi_k(t') \succeq \pi_k(\pi_k(t)) = \pi_k(t)$, by **iii**). Hence $t' \prec \pi_k(t)$ iff $\pi_k(t') \prec \pi_k(t)$. This establishes (5.19). Concerning (5.20), we have seen above (see proof of **iv**)) that $\{t' \in \mathbb{N}^d : \pi_k(t) = \pi_k(t')\} = \mathcal{C}_{k,\pi_k(t)}$, therefore

$$\Pi_k(t) = \Pi_k^\prec(t) \cup \{t' \in \mathbb{N}^d : \pi_k(t) = \pi_k(t')\} = \Pi_k^\prec(t) \cup \mathcal{C}_{k,\pi_k(t)}.$$

vi): It holds $\pi_{k-1}(t) \succeq \pi_k(t)$ from **ii**). Then from (5.19), we deduce the inclusion $\Pi_k^\prec(t) \subseteq \Pi_{k-1}^\prec(t)$. The inclusion $\Pi_{k-1}^\prec(t) \subseteq \Pi_{k-1}(t)$ is immediate from the definitions (5.17), (5.18). Finally, for any $t' \in \Pi_{k-1}(t)$, by definition $\pi_{k-1}(t') \preceq \pi_{k-1}(t)$, so by **iii**) and $\pi_k \circ \pi_{k-1} = \pi_k$, it holds $\pi_k(t') \preceq \pi_k(t)$, hence $t' \in \Pi_k(b)$, proving the last inclusion. □

Remark 5.5.2. Choice of the lexicographical order in the definition (5.16) is largely arbitrary; any total order on $\mathbb{N}^d$ that is compatible with the coordinate-wise partial order would work, since it would result in the same properties as above, which are the only ones we will be using in the sequel.

For $t \in \mathbb{N}^d$ and an integer k, define

$$\mathcal{F}_k^\prec(t) := \mathfrak{S}(X_{t'}, t' \in \Pi_k^\prec(t)), \qquad \mathcal{F}_k(t) := \mathfrak{S}(X_{t'}, t' \in \Pi_k(t)), \tag{5.22}$$

where $\Pi_k^\prec, \Pi_k$ are as defined in (5.17), (5.18) (and $\mathfrak{S}(\emptyset)$ is the trivial σ-algebra).
For every element $t \in \mathbb{N}^d$, using the fact that $\Pi_k^\prec(t) = \emptyset$ for $k > \log_2 t$, we write the decomposition

$$X_t - \mathbb{E}[X_t] = \Big(X_t - \mathbb{E}\big[X_t | \mathcal{F}_0^\prec(t)\big]\Big) + \sum_{k=1}^{\lfloor \log_2 \|t\|_\infty \rfloor + 1} \big(\mathbb{E}\big[X_t | \mathcal{F}_k^\prec(t)\big] - \mathbb{E}\big[X_t | \mathcal{F}_{k-1}^\prec(t)\big]\big).$$

For any finite subset $A \subset \mathbb{N}^d$, denoting $\pi_k(A) = \{\pi_k(t), t \in A\} \subset 2^k\mathbb{N}^d$ and $\|A\|_\infty = \max_{t \in A} \|t\|_\infty$,

we have $A = \uplus_{b\in\pi_k(A)}(A\cap \mathcal{C}_{k,b})$, hence:

$$\begin{aligned} S_A &= \sum_{t\in A}(X_t - \mathbb{E}[X_t]) \\ &= \sum_{t\in A}\Big(X_t - \mathbb{E}\big[X_t|\mathcal{F}_0^{\prec}(t)\big]\Big) + \sum_{k=1}^{\lfloor\log_2\|A\|_\infty\rfloor+1}\sum_{b\in\pi_k(A)}\sum_{t\in\mathcal{C}_{k,b}\cap A}\big(\mathbb{E}\big[X_t|\mathcal{F}_k^{\prec}(t)\big] - \mathbb{E}\big[X_t|\mathcal{F}_{k-1}^{\prec}(t)\big]\big) \\ &= \sum_{k=0}^{\lfloor\log_2\|A\|_\infty\rfloor+1}\sum_{b\in\pi_k(A)} Z_{b,k}(A), \end{aligned} \tag{5.23}$$

where

$$Z_{t,0}(A) := X_t - \mathbb{E}\big[X_t|\mathcal{F}_0^{\prec}(t)\big]; \tag{5.24}$$

$$\text{and for } k\geq 1: \quad Z_{b,k}(A) := \sum_{t\in\mathcal{C}_{k,b}\cap A}\big(\mathbb{E}\big[X_t|\mathcal{F}_k^{\prec}(t)\big] - \mathbb{E}\big[X_t|\mathcal{F}_{k-1}^{\prec}(t)\big]\big). \tag{5.25}$$

Lemma 5.5.3. *Let $A\subset\mathbb{N}^d$ be a finite set. Let k be a fixed integer. Then $(Z_{b,k}(A),\mathcal{F}_k(b))_{b\in\pi_k(A)}$ is a martingale difference, where $\pi_k(A)$ is ordered by the total order $\preceq$ defined by (5.16).*

Proof We start with the special case $k=0$. In this case, since $\pi_0(t)=t$, we have $\mathcal{F}_0^{\prec}(t)=\mathfrak{S}(X_{t'},t'\prec t)$, and $\mathcal{F}_0(t)=\mathfrak{S}(X_{t'},t'\preceq t)$. It is straightforward that $Z_{0,t}(A)$ is $\mathcal{F}_0(t)$-measurable, and that for any $t'\prec t$ we have $\mathcal{F}_0(t')\subseteq\mathcal{F}_0^{\prec}(t)$ thus $\mathbb{E}[Z_{0,t}(A)|\mathcal{F}_{t'}]=0$; hence the claim. Let $k\geq 1$ be a fixed integer. The claim for $(Z_{b,k})$ relies on points (4) and (6) of Lemma 5.5.1, which straightforwardly implies for any $t\in\mathcal{C}_{k,b}$ that $\mathcal{F}_k^{\prec}(b)=\mathcal{F}_k^{\prec}(t)\subseteq\mathcal{F}_{k-1}^{\prec}(t)\subseteq\mathcal{F}_k(t)=\mathcal{F}_k(b)$. Thus, $Z_{b,k}(A)$ is $\mathcal{F}_k(b)$-measurable, and for any $b'\prec b$, since $\mathcal{F}_{k,b'}\subseteq\mathcal{F}_{k,b}^{\prec}$, it holds $\mathbb{E}\big[Z_{k,b}|\mathcal{F}_{k,b'}\big]=0$, implying the claim. □

Notice that our proof relies on multi-scale martingale decomposition which is used to obtain control of the set $\|S_A\|_p$ from the control of the martingale increments $\|Z_{b,k}\|_p$ using Burkhölder's inequality (or, analogously, using subgaussian norm from the Azuma's inequality). However, while the control obtained this way is optimal for segments in dimension 1 (see Peligrad et al. (2006)), it turns out that it is not the case for rectangles in dimension $d\geq 2$. The reason is that there are "too many" elements in the sum over the cell $\mathcal{C}_{k,b}$, appearing in the martingale definition (5.25), that are close to the boundary of the cell and thus to $\Pi_k^{\prec}(b)$, preventing the efficient usage of the weak dependence assumption (1).

To alleviate this issue we exclude from the sum the elements which are close to boundaries of the cell on every scale (this is described in the set $\mathfrak{F}_\delta$). The remaining elements are then sufficiently "separated" from the boundaries. We recall that $\Lambda_{k,\delta} := 2^k\mathbb{N}_{k>0}+\delta$. Firstly we need the following supporting result which estimates the distance from any element from the set t to the boundary of the neighbor cell.

Lemma 5.5.4. *For any $k\in\mathbb{N}$, $\delta\in\{2^k\}_0$ and $t\in(\mathbb{N}\setminus\Lambda_{k,\delta})^d$, it holds*

$$d_\infty(t,\Pi_k^{\prec}(t))\geq\delta+1.$$

Proof **i**) from Lemma 5.5.1 implies that $\preceq$ is compatible with the partial coordinate-wise order $\leq_{\mathrm{cw}}$. This implies in particular that any t' such that $\pi_k(t)\leq_{\mathrm{cw}}t'$ satisfies $\pi_k(\pi_k(t))=\pi_k(t)\preceq\pi_k(t')$, and thus cannot belong to $\Pi_k^{\prec}(t)$. Therefore, for any $t'\in\Pi_k^{\prec}(t)$, there exists a coordinate i such that $t'_i<\pi_k(t_i)$. In particular, $\pi_k(t_i)>0$, hence $\pi_k(t_i)\in 2^k\mathbb{N}_{>0}$. On the other hand, if we assume $t\in(\mathbb{N}\setminus\Lambda_{k,\delta})^d$ then $t_i\in\mathbb{N}\setminus(2^k\mathbb{N}_{>0}+\{\delta\}_0)$. Since $t'_i<\pi_k(t_i)\leq t_i$, it must hold $t_i-t'_i\geq t_i-\pi_k(t_i)+1\geq\delta+1$, implying the claim. □

Proposition 5.5.5. *Let $\mathcal{R}=\prod_{i=1}^d\{N_i\}_0$ be a d-dimensional rectangle of side-lengths $N_i\geq 1$, $i=1,\dots,d$, and $m(\mathcal{R}):=\max_{i=1,\dots,d}\lfloor\log_2 N_i\rfloor$. Let $\boldsymbol{\delta}=(\delta_k)_{k\geq 1}$ be a fixed decreasing sequence of*

integers with $\delta_k \leq 2^k$, $k \geq 1$, and put $\delta_0 = 0$ and $\mathfrak{F}_{\boldsymbol{\delta}}$ be as defined in (5.6). If process $(X_t)_{t\in\mathbb{N}^d}$ satisfies weak-dependency assumption 1 with $2 \leq p < \infty$ and some rate $\varphi_p(\cdot)$ then it holds

$$\|S_{\mathcal{R}\cap\mathfrak{F}_{\boldsymbol{\delta}}}\|_p \leq \frac{1}{2} C_p \Psi_p(\boldsymbol{\delta}, \mathcal{R}), \tag{5.26}$$

where $C_p := 4\sqrt{p}$; and if assumption 1 is satisfied for $p = \infty$, then it holds

$$\|S_{\mathcal{R}\cap\mathfrak{F}_{\boldsymbol{\delta}}}\|_{SG} \leq \frac{1}{2} C_\infty \Psi_p(\boldsymbol{\delta}, \mathcal{R}), \tag{5.27}$$

where $C_\infty := 10$, $\|X\|_{SG} := \inf_{c>0}\{\mathbb{E}[\exp(\lambda X)] \leq \exp\Big(\frac{\lambda^2 c^2}{2}\Big), \}$ subgaussian norm and

$$\Psi_p(\boldsymbol{\delta}, \mathcal{R}) = 2M_p\sqrt{|\mathcal{R}|}\Big(1 + \varphi_p(1) + \sum_{k=1}^{m(\mathcal{R})+1} \varphi_p\big(\delta_{k-1}+1\big)\sqrt{|\mathcal{C}_{k,0}\cap\mathcal{R}|}\Big).$$

Proof We use the decomposition (5.23) with $A = \mathcal{R}\cap\mathfrak{F}_\delta$, so that by the triangle inequality

$$\|S_{\mathcal{R}\cap\mathfrak{F}_{\boldsymbol{\delta}}}\|_p \leq \sum_{k=0}^{m(\mathcal{R})+1} \Big\| \sum_{b\in\pi_k(\mathcal{R}\cap\mathfrak{F}_{\boldsymbol{\delta}})} Z_{b,k}(\mathcal{R}\cap\mathfrak{F}_{\boldsymbol{\delta}}) \Big\|_p, \tag{5.28}$$

where $Z_{b,k}(\mathcal{R}\cap\mathfrak{F}_{\boldsymbol{\delta}})$ is defined in (5.24), (5.25). We now estimate the norm of the martingale increments $Z_{b,k}(\mathcal{R}\cap\mathfrak{F}_{\boldsymbol{\delta}})$ using Assumption 1. We will denote below $Z_{b,k} = Z_{b,k}(\mathcal{R}\cap\mathfrak{F}_{\boldsymbol{\delta}})$ and $S_k = \pi_k(\mathcal{R}\cap\mathfrak{F}_{\boldsymbol{\delta}})$ for ease of notation. As a direct consequence of Lemma 5.5.4, for any $t \in \mathfrak{F}_{\boldsymbol{\delta}}$ it holds $\mathcal{F}_k^{\prec}(t) \subset \mathcal{M}_{t,\delta_k+1}$ (as defined in Assumption 1). Therefore, for $k = 0$,

$$\begin{aligned}
\|Z_{b,0}\|_p &= \big\|X_b - \mathbb{E}\big[X_b|\mathcal{F}_0^{\prec}(b)\big]\big\|_p \\
&\leq \big\|X_b - \mathbb{E}\big[X_b|\mathcal{M}_{b,1}\big]\big\|_p \\
&\leq \|X_b - \mathbb{E}[X_b]\|_p + \big\|\mathbb{E}\big[X_b|\mathcal{M}_{b,1}\big] - \mathbb{E}[X_b]\big\|_p \\
&\leq M_p(\varphi(0) + \varphi(1)),
\end{aligned} \tag{5.29}$$

while for $k \geq 1$:

$$\begin{aligned}
\|Z_{b,k}\|_p &= \Big\| \sum_{t\in\mathcal{C}_{k,b}\cap\mathcal{R}\cap\mathfrak{F}_{\boldsymbol{\delta}}} \big(\mathbb{E}\big[X_t|\mathcal{F}_k^{\prec}(t)\big] - \mathbb{E}\big[X_t|\mathcal{F}_{k-1}^{\prec}(t)\big]\big)\Big\|_p \\
&\leq \sum_{t\in\mathcal{C}_{k,b}\cap\mathcal{R}\cap\mathfrak{F}_{\boldsymbol{\delta}}} \Big(\Big\|\mathbb{E}\big[X_t|\mathcal{F}_k^{\prec}(t)\big] - \mathbb{E}[X_t]\Big\|_p + \Big\|\mathbb{E}\big[X_t|\mathcal{F}_{k-1}^{\prec}(t)\big] - \mathbb{E}[X_t]\Big\|_p\Big) \\
&\leq \sum_{t\in\mathcal{C}_{k,b}\cap\mathcal{R}\cap\mathfrak{F}_{\boldsymbol{\delta}}} \Big(\Big\|\mathbb{E}\big[X_t|\mathcal{M}_{t,\delta_k+1}\big] - \mathbb{E}[X_t]\Big\|_p + \Big\|\mathbb{E}\big[X_t|\mathcal{M}_{t,\delta_{k-1}+1}\big] - \mathbb{E}[X_t]\Big\|_p\Big) \\
&\leq |\mathcal{C}_{k,b}\cap\mathcal{R}| M_p\big(\varphi(\delta_k+1) + \varphi(\delta_{k-1}+1)\big).
\end{aligned} \tag{5.30}$$

Note that we can subsume (5.29) into (5.30) by putting formally $\delta_{-1} := -1$. Since, by Lemma 5.5.3, the sequence $(Z_{b,k}, \mathcal{F}_k(b))_{b\in S_k}$ is a martingale difference sequence over b for fixed k, if $p \in [2,\infty)$ we can

apply Burkholder's inequality (Burkholder (1966)). We obtain, after combining with the above estimate:

$$\begin{aligned}
\Big\| \sum_{b\in C_{k,b}\cap\mathcal{R}} Z_{b,k} \Big\|_p &\le \sqrt{p} \Big\| \Big(\sum_{b\in C_{k,b}\cap\mathcal{R}} Z_{b,k}^2 \Big)^{\frac{1}{2}} \Big\|_p \\
&= \sqrt{p} \Big\| \sum_{b\in C_{k,b}\cap\mathcal{R}} Z_{b,k}^2 \Big\|_{p/2}^{\frac{1}{2}} \\
&\le \sqrt{p} \Big(\sum_{b\in S_k} \|Z_{b,k}\|_p^2 \Big)^{\frac{1}{2}} \\
&\le \sqrt{p} M_p \big(\varphi(\delta_k+1)+\varphi(\delta_{k-1}+1)\big) \Big(\sum_{b\in\pi_k(\mathcal{R})} |\mathcal{C}_{k,b}\cap\mathcal{R}|^2 \Big)^{\frac{1}{2}} \\
&\le 2\sqrt{p} M_p \varphi_p(\delta_{k-1}+1) \Big(\sum_{b\in\pi_k(\mathcal{R})} |\mathcal{C}_{k,b}\cap\mathcal{R}|^2 \Big)^{\frac{1}{2}} \qquad (5.31)
\end{aligned}$$

We now concentrate on the estimate for $\sum_{b\in\pi_k(\mathcal{R})}|\mathcal{C}_{k,b}\cap\mathcal{R}|^2$. Put $q_i := \lfloor \frac{N_i}{2^k} \rfloor$ and $r_i := N_i - q_i 2^k$, for $i = 1,\dots,d$. Observe that $\pi_k(\mathcal{R}) = \prod_{i=1}^d (2^k\{q_i+1\}_0)$; for $b = (b_1,\dots,b_d) \in \pi_k(\mathcal{R})$, the set $\mathcal{C}_{k,b}\cap\mathcal{R}$ is a hyperrectangle with side-lengths:

$$\ell_k(b_i,N_i) := \begin{cases} 2^k & \text{if } b_i < 2^k q_i; \\ r_i & \text{if } b_i = 2^k q_i. \end{cases}$$

Hence, it holds

$$\begin{aligned}
\sum_{b\in\pi_k(\mathcal{R})} |\mathcal{C}_{k,b}\cap\mathcal{R}|^2 = \sum_{b\in\prod_{i=1}^d (2^k\{q_i+1\}_0)} \prod_{i=1}^d \ell_k(b_i,N_i)^2 &= \prod_{i=1}^d \sum_{j=0}^{q_i} \ell_k(2^k j, N_i)^2 \\
&= \prod_{i=1}^d (q_i 2^{2k} + r_i^2) \\
&\le \prod_{i=1}^d \big((N_i - r_i)\min(2^k,N_i) + r_i \min(2^k,N_i) \big) \\
&= \prod_{i=1}^d \big(N_i \min(2^k,N_i) \big) \\
&= |R||R\cap\mathcal{C}_{0,k}|. \qquad (5.32)
\end{aligned}$$

The claimed estimate for $2 \le p < \infty$ follows by using (5.31) and (5.32) into (5.28) and straightforward computations. In the case of $p = \infty$, we can apply the bounded martingale difference inequality (Azuma (1967)) stating that the sum $\sum_{b\in S_k} Z_{b,k}$ is sub-Gaussian such that $\Big\|\sum_{b\in S_k} Z_{b,k}\Big\|_{SG} \le \Big(\Big\|\sum_{b\in S_k} Z_{b,k}^2\Big\|_\infty\Big)^{\frac{1}{2}}$ and using triangle inequality for the sub-gaussian norm

$$\Big\| \sum_{k=0}^{m(\mathcal{R})+1} \sum_{b\in\pi_k(\mathcal{R}\cap\mathfrak{F}_\delta)} Z_{b,k}(\mathcal{R}\cap\mathfrak{F}_\delta) \Big\|_{SG}$$

over scales $k \in \{m(\mathcal{R})+2\}_0$. All other arguments are as in the case $p < \infty$. □

The elements from the set $\mathcal{R} \cap \mathfrak{F}_{\boldsymbol{\delta}}^c$ form a multi-scale "frame" of sufficiently small cardinality which will be handled by recursive induction. To justify the usage of induction we prove the following general statement that the random field indexed with the set of excluded elements satisfies the weak-dependence Assumption (5.1). We recall that if $A \subset \mathbb{N}$, and $j \in \{|A|\}_0$ so $j : \mathbf{A}$ denotes the $(j+1)$-th element of $\mathbf{A}$ in the increasing order; we used $t : \mathbf{A} := (t_1 : A_1, \ldots, t_d : A_d)$, $\mathcal{K}(\mathbf{A}) = \prod_{i=1}^d \{|A_i|\}_0$ and

$$\widetilde{X}_t^{(\mathbf{A})} = X_{t:\mathbf{A}}, \qquad t \in \mathcal{K}(\mathbf{A}) \qquad \widetilde{S}_{\mathcal{K}(\mathbf{A})}^{(\mathbf{A})} := \sum_{t \in \mathcal{K}(\mathbf{A})} \widetilde{X}_t^{(\mathbf{A})} = S_{\mathbf{A}}.$$

The following results states that if the original process $(X_t)_{t\in\mathbb{N}^d}$ satisfies Assumption 1, then so does $(\widetilde{X}_t^{\mathbf{A}})_{t\in\mathbb{N}^d}$ (defined as above, then padded with zeros for $t \notin \mathcal{K}(\mathbf{A})$).

Lemma 5.5.6. *Let $2 \le p \le \infty$ and the process $(X_t)_{t\in\mathbb{N}^d}$ satisfies the weak-dependency assumption 5.1 with rate $\varphi_p(\cdot)$ and $\mathbf{A} \subset \mathbb{N}^d$. Then the "compressed" version $\left(\widetilde{X}_t^{\mathbf{A}}\right)_{t\in\mathbb{N}^d}$ as defined by (5.8) satisfies the weak-dependency assumption 5.1 with the same rate $\varphi_p(\cdot)$.*

Proof Since for every $t \notin \mathcal{K}(\mathbf{A})$ we have by construction that $\tilde{X}_t^{(\mathbf{A})} = 0$, the claim evidently follows from the definition of weak-dependency assumption 5.1. It is therefore sufficient to prove the property only for elements $t \in \mathcal{K}(\mathbf{A})$. For every $t \in \mathcal{K}(\mathbf{A})$ for every $u \in \mathcal{K}(\mathbf{A})$ since $\|u : \mathbf{A} - t : \mathbf{A}\|_{d,\infty} \ge \|u - t\|_{d,\infty}$, by Jensen's inequality and using Assumption 5.1 we have:

$$\begin{aligned}
\left\|\mathbb{E}\left[\widetilde{X}_t^{(\mathbf{A})}|\widetilde{\mathcal{M}}_{t,r}\right] - \mathbb{E}\left[\widetilde{X}_t^{(\mathbf{A})}\right]\right\|_p &= \left\|\mathbb{E}\left[X_{t:\mathbf{A}}|\sigma\Big(X_{u:\mathbf{A}} : \|u-t\|_{d,\infty} \ge r\Big)\right] - \mathbb{E}[X_{t:\mathbf{A}}]\right\|_p \\
&= \left\|\mathbb{E}\left[\mathbb{E}\left[X_{t:\mathbf{A}}|\sigma\big(X_{u:\mathbf{A}} : \|u:\mathbf{A}-t:\mathbf{A}\|_{d,\infty} \ge r\big) - \mathbb{E}[X_{t:\mathbf{A}}]\right]|\sigma\Big(X_{u:\mathbf{A}} : \|u-t\|_{d,\infty} \ge r\Big)\right]\right\|_p \\
&\le \left\|\mathbb{E}\left[X_{t:\mathbf{A}}|\sigma\big(X_{u:\mathbf{A}} : \|u:\mathbf{A}-t:\mathbf{A}\|_{d,\infty} \ge r\big) - \mathbb{E}[X_{t:\mathbf{A}}]\right]\right\|_p \\
&\le \left\|\mathbb{E}\left[X_t|\sigma\big(X_u : \|u-t\|_{d,\infty} \ge r\big) - \mathbb{E}[X_t]\right]\right\|_p \le M_p\varphi(r).
\end{aligned}$$

□

Bibliography

Adams, H. and Fournier, J. (2003). *Sobolev spaces*. Academic Press.

Agarwahl, A. and Duchi, J. (2012). The generalization ability of online algorithms for dependent data. *IEEE Transactions on Information Theory*, 59(1):573–587.

Amat, C., Michalski, T., and Stoltz, G. (2018). Fundamentals and exchange rate forecastability with simple machine learning methods. *Journal of International Money and Finance*, 88:1–24.

Anantharam, V., Varaiya, P., and Walrand, J. (1987). Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: I.i.d. rewards. *IEEE Transactions on Automatic Control*, 32(11):968–977.

Andrews, D. (1984). Non-strong mixing autoregressive processes. *J.Appl. Prob.*, 21:930–934.

Andrews, D. (1988). Laws of large numbers for independent non identically distributed random variables. *Econometric Theory*, 4:458–467.

Argyriou, A. and Dinuzzo, F. (2014). A unifying view of representer theorems. In Xing, E. P. and Jebara, T., editors, *International Conference on Machine Learning 31 (ICML 2014)*, volume 32 of *Proceedings of Machine Learning Research*, pages 748–756.

Audibert, J. and Bubeck, S. (2009). Minimax policies for adversarial and stochastic bandits. In *COLT*.

Audiffren, J. and Ralaivola, L. (2015). Cornering stationary and restless mixing bandits wuth remix-ucb. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pages 3339–3347.

Auer, P. (2002). Using confidence bounds for exploration-explotation trade-offs. *Journal of Machine Learning Research*, 3:397–422.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. (2002). The nonstochastic multiarmed bandit problem. *SIAM Journal of Computing*, 32(1):48–77.

Auer, P. and Ortner, R. (2010). Ucb revisited: improved upper bounds for the stochastic multi-armed bandit problem. *Peroodica Mathematica Hungarica*, 61(1-2):55–65.

Azoury, K. and Warmuth, M. (2001). Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine learning*, 43:211–246.

Azuma, K. (1967). Weighted sums of certain dependent variables. *Tohoku Mathematical Journal*, 19(3):357–367.

Bauer, F., Pereverzev, S., and Rosasco, L. (2009). On regularization algorithms in learning theory. *Journal of Complexity*, 1(23):53–72.

Benett, K. and Bredensteiner, J. (2000). Duality and geometry in support vector machine classifiers. In Langley, P., editor, *International Conference on Machine Learning 17 (ICML 2000)*, pages 57–64.

Bernstein, S. (1924). On a modification of chebyschev's inequality and of the error formula of laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math*, 4(5).

Bhatia, R. (1997). *Matrix analysis*. Springer.

Bickel, P. and Buehlmann, P. (1999). A new mixing motivation and functional central limit theorems for a sieve bootstrap in times series. *Bernoulli*, (5):413–446.

Biermé, H. and Durieu, O. (2014). Invariance principles for self-similar set-indexed random fields. *Trans. Amer. Math Soc.*, 366(11):5963–5989.

Bishop, C. (2006). *Pattern Recongnition and Machine Learning*. Springer.

Blackwell, A. (1956). An analog of minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6(1):1–8.

Blanchard, G., Bousquet, O., and Zwald, L. (2007). Statistical properties of kernel principal component analysis.hal hal-00373789. *Machine Learning*, 3:259–294.

Blanchard, G., Lee, G., and Scott, C. (2011). Generalizing from several related classification tasks to a new unlabeled sample. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Inf. Proc. Systems 24 (NIPS 2011)*, pages 2438–2446.

Blanchard, G. and Mücke, N. (2018). Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 18(4):971–1013.

Blanchard, G. and Zadorozhnyi, O. (2019). Concentration of weakly-dependent Banach-valued sums and applications to statistical learning methods. *Bernoulli*, 25(4B):3421–3458.

Bosq, D. (2000). *Linear Processes in Function Spaces: Theory and Applications*, volume 149 of *Lecture Notes in Statistics*. Springer.

Bosq, P. (1993). Bernstein-type large deviations inequalities for partial sums of strong mixing processes. *Statistics*, 24(1):59–70.

Bradley, R. (2002). *Introduction to strong mixing coefficients*. Technical report, I.U. Bloomington.

Bradley, R. (2005). Basic properties of strong mixing conditions. *Probability Surveys*, 2:107–144.

Bradley, R. (2007). *Introduction to Strong Mixing Conditions*, volume 1,2,3. Kendrick Press.

Bradley, R., Bryc, W., and Janson, S. (1987). Remarks on the foundations of measures of dependence. *New Perspectives in Theoretical and Applied Statistics*, pages 421–437.

Brezis, H. and Mironescu, P. (2018). Gagliardo-nierenberg inequalities and non-inequalities. *Annales de l'Institut de Henri Poincare*, 1:1355–1376.

Bubeck, S. (2010). *Bandit Games and Clustering Foundations*. PhD thesis, Universite des Sciences et technologie de Lille - Lille I.

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122.

Bubeck, S., Cesa-Bianchi, N., and Lugosi, G. (2013). Bandits with heavy tail. *IEEE Transactions on Information Theory*.

Bubeck, S., Munos, R., and Stoltz, G. (2009). Pure exploration in multi-armed bandits problems. In *In Algorithmic Learning Theory*, pages 23–37. Springer.

Burkholder, D. (1966). Martingale transforms. *Annals of Mathematical Statistics*, 37(6):1494–1504.

Cairoli, R. (1969). Un théoréme de convergence pour martingales a indices multiples. *C.R.Acad.Sci. Paris*, 269:587–589.

Canda, K. (1974). Strong mixing properties of linear stochastic processes. *Journal of Applied Pobability*, 11:401–408.

Canu, S., Mary, X., and Rakotomamonjy, A. (2003). *Functional learning through kernel*, chapter 5, pages 89–110. IOS Press.

Caponetto, A. and De Vito, E. (2005). Risk bounds for regularized least squares algorithm with operator-valued kernels. *Technical report, Massachusetts Institute of Technology, Cambridge*, CBCL Paper 249(015).

Caponetto, A. and De Vito, E. (2007). Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368.

Caponetto, A. and E.De.Vito (2006). Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, pages 331–368.

Cesa-Bianchi, N. (1999a). Analysis of two gradient-based algorithms for online regression. *Journal of Computer and System Sciences*, 59:392–411.

Cesa-Bianchi, N. (1999b). Analysis of two gradient-based algorithms for online regression. *Journal Computational System Sci.*, pages 392–411.

Cesa-Bianchi, N., Freund, Y., Haussler, D., and Helmbold, D. Schapire, R. W. M. (1997). How to use expert advice. *Journal of the ACM*, 44(3):427–485.

Cesa-Bianchi, N., Gaillard, P., Gentile, C., and Gerchinovitz, S. (2017). Algorithmic chaining and the role of partial feedback in online nonparametric learning. *arXiv preprint arXiv:1702.08211*.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning and Games*. Cambridge University Press.

Christmann, A. and Steinwart, I. (2007). Consistency and robustness of kernel-based regression in convex-risk minimization. *Bernoulli*, 3(13):799–819.

Clarkson, J. (1936). Uniformly convex spaces. *Transactions of the American Mathematical Society*, 40:396–414.

Combettes, P. L., Salzo, S., and Villa, S. (2018). Regularized learning schemes in feature banach spaces. *Analysis and Applications*, 16(01):1–54.

Cortez, P. and Morais, A. (2007). A data mining approach to predict forest fires using meteorological data. Online dataset.

De Vito, E., Rosasco, L., and Caponnetto, A. (2006). Discretization error analysis for tikhonov regularization. *Analysis and Applications*, 4(1):81–99.

Dedecker, J. (1991). Exponential inequalities and functional central limit theorems for random fields. *ESAIM Probability and Statistics*, 5:77–104.

Dedecker, J., Doukhan, P., Lang, G., Leon, R., Louhichi, S., and Prieur, C. (2007). *Weak dependence with examples and applications*. Springer, New York.

Dedecker, J. and Merlevede, F. (2015). Moment bounds for dependent sequences in smooth banach spaces. *Stochastic Processes and their Applications*, 125(9):3401–3429.

Dedecker, J. and Prieur, C. (2005). New dependence coefficients. examples and application to statistics. *Prob Theory Relatex Fields*, 2(132):203–236.

Dehling, H. and Philipp, W. (1982). Almost sure invariance principles for weakly dependent vector-valued random variables. *Annals of probability*, 10:689–701.

Desautels, T., Krause, A., and Burdick, J. W. (2014). Parallelizing exploration-exploitation tradeoffs in gaussian process bandit optimization. *Journal of Machine Learning Research*, pages 4053–4103.

Devaine, M., Gaillard, P., Goude, Y., and Stoltz, G. (2013). Forecasting electricity consumption by aggregating specialized experts - a review of the sequential aggregation of specialized experts, with an application to slovakian and french country-wide one-day-ahead (half-)hourly predictions. *Machine Learning*, 90(2):231–260.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A probabilistic theory of pattern recognition*. Springer, New York.

Doukhan, P. (1994). *Mixing: properties and examples*. Springer, Berlin, lecture notes in statistics edition.

Doukhan, P., Léon, J., and Portal, F. (1984). Vitesse de convergence dans le théorème central limit pour des variables aléatoires mélangeantes à valeurs dans un espace de hilbert. *C.R. Acad. Sci. Paris*, pages 305–308.

Doukhan, P. and Louhichi, S. (1999). A new weak dependence condition and application to moment inequalities. *Stoch. Proc. Appl.*, 84:313–342.

Eberts, M. and Steinwart, I. (2011). Optimal learning rates for least squares svms using gaussian kernels. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 24, pages 1539–1547. Curran Associates, Inc.

Edmunds, D. and Triebel, H. (1996). *Function Spaces, Entropy Numbers,Differential Operators*. Cambridge University Press.

El Machkouri, M., Volny, D., and Wu, W. (2013). A central limit theorem for stationary random fields. *Stochastic Processes and their Applications*, 123:1–14.

Engl, H., Hanke, M., and Neubauer, A. (1996). Regularization of inverse problems. *Mathematics and its Applications*, 375.

Engl, H., Hanke, M., and Neubauer, A. (2000). *Regularization of inverse problems*. Springer Netherlands.

Esary, J., Proschan, F., and Walkup, D. (1985). Association of random variables with applications. *Ann. Math. Statist.*, 38:1466–1476.

Evan-Dar, E., Mannor, S., and Mansour, Y. (2006). Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7:1079–1105.

Even-Dar, E., Mannor, S., and Mansour, Y. (2002). Pac bounds for multi-armed bandits and markov decision processes. In *In Computational Learning Theory*, pages 255–270. Springer.

Fan, X., Grama, I., and Liu, Q. (2015). Exponential inequalities for martingales with applications. *Electron. J. Probab.*, 20(1):1–22.

Fazekas, I. (1983). Convergence of vector-valued martingales with multidimensional indices. *Publ. Math Debrecen*, 30:157–164.

Fazekas, I. (2005). Burkholder's inequality for multiindex martingales. *Annales Mathematicae et Informaticae*, 32:45–51.

Fischer, S. and Steinwart, I. (2017). Sobolev norm learning rates for regularized least-squares algorithms. *Arxiv*, pages 1–26.

Fortuyn, C., Kastelyan, P., and Ginibre, J. (1971). Correlation inequalities in some partially ordered sets. *Com. Math. Phys.*, 74:119–122.

Foster, D. (1991). Prediction in the worst case. *Annals of Statistics*, 19:1084–1090.

Foster, D. and Vohra, R. (1997). Calibrated learning and correlated equilibrium. *Games and Economic behaviour*, 21(1):40–55.

Freedman, D. (1975). On tail probabilities for martingales. *Ann. Probab.*, 3(1):100–118.

Gaillard, P. and Gerchinovitz, S. (2015). A chaining algorithm for online nonparametric regression. In *Proceedings of The 28th Conference on Learning Theory*, volume 40, pages 764–796.

Gaillard, P., Gerchinovitz, S., Huard, M., and Stoltz, G. (2019). Uniform regret bounds over $\mathbb{R}^d$ for the sequential linear regression problem with the square loss. In Garivier, A. and Kale, S., editors, *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pages 404–432, Chicago, Illinois. PMLR.

Gammerman, A., Kalnishkan, Y., and Vovk, V. (2004). On-line prediction with kernels and the complexity approximation principle. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 170–176.

Garivier, A. and Cappé, O. (2011). The kl-ucb algorithm for bounded stochastic bandits and beyond. In Sham Kakade, U. L., editor, *Proceedings of the 24th Conference on Learning Theory*, volume 19, pages 359–376.

Gerchinovitz, S. (2013). Sparsity regret bounds for individual sequences in online linear regression. *Journal of Machine Learning Research*, pages 729–769.

Gerchinovitz, S. and Yu, J. (2013). Adaptive and online linear regression on ℓ_1 balls. *Theoretical Computer Science*.

Gerchinowitz, S. and Lattimore, T. (2016). Refined lower bounds for adversarial bandits. In *In Proceedings of Advances in Neural Information Processing Systems*.

Giraudo, D. (2018). Invaraince principle via orthomartingale approximation. *Stochastics and Dynamics*, 18(6).

Giraudo, D. (2019). Deviation inequalities for banach space valued martingales differences sequences and random fields. *ESAIM: Probability and Statistics*, 23:922–946.

Giraudo, D. (2020). Bound on the maximal function associated to the law of iterated logarithms for bernoulli random fields. *Preprint*.

Grünewälder, S. and Khaleghi, A. (2017). Approximations of the restless bandit problem. *Preprint*.

Guha, S., Munagala, K., and M., P. (2010a). Multi-armed bandit problems with delayed feedback. Arxiv. Preprint.

Guha, S., Munagala, K., and Shi, P. (2010b). Approximation algorithms for restless bandit problems. *Journal of the ACM (JACM)*, 58(1).

Ǵyörfi, L. (2002). *A Distribution-Free theory of nonparametric regression*. Springer.

Hang, H. and Steinwart, I. (2017). A bernstein-type inequality for some mixing processes and dynamical systems with application to learning. *Ann. Statist.*, 45(2):708–743.

Hannan, E. (1973). Central limit theorems for time series regression. *Z. Wahrscheinlichkeitstheor. Verwandte Geb.*, 26:157–170.

Hannan, J. (1957). Approximations to bayes risk in repeated play. *Contributions to the Theory of Games*, (3):97–139.

Hart, S. and Mas-Colell, A. (2000). A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150.

Hazan, E.and Kalai, A. and Kale, S.and Agarwal, A. (2006). Logarithmic regret algorithms for online convex optimization. *COLT*.

Hazan, E. (2006). *Efficient Algorithms for Online Convex Optimization and Their Applications*. PhD thesis, Princeton University, USA.

Hein, M., Bousquet, O., and Scholköpf, B. (2005). Maximal margin classification for metric spaces. *Journal of Computer and System Sciences*, 71:333–359.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.

Horváth, L. and Kokoszka, P. (2012). *Inference for functional data with applications*, volume 200 of *Springer series in Statistics*. Springer.

Ibragimov, I. (1959). Some limit theorems for stochastic processes stationary in the strict sense. *Dolk. Akad. Nauk. USSSR*, (125):711–714.

Jarner, S. and Roberts, G. (2002). Polynomial convergence rates of markov chains. *The Annals of Applied Probability*, 12(1):224–247.

Jézéquel, R., Gaillard, P., and Rudi, A. (2019). Efficient online learning with kernels for adversarial large scale problems. In *Advances in Neural Information Processing Systems*, pages 9427–9436.

Jirak, M. (2016). Berry-essen theorems under weak-dependence. *Ann. Probab*, 44:2024–2063.

Jirak, M. (2018). Rate of convergence for hilbert space valued processes. *Bernoulli*. (to appear).

Joulani, P., Gyorgy, A., and C., S. (2013). Online learning under delayed feedback. In *ICML 2013*, volume 28, pages 1453–1461.

Kallenberg, O. (2017). *Random measures. Theory and Applications. 2nd edition.* Springer.

Kimeldorf, G. and Wahba, G. (1970). A correspondence between bayesian estimation of stochastic processes and smoothing by splines. *Ann. Math. Stat.*, 41(495–502).

Kivinen, J. and Warmuth, M. K. (1997). Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132:1–63.

Klenke, A. (2010). *Probability Theory*. Springer, New York, second edition edition.

Klicnarová, J., Volny, D., and Wang, Y. (2016). Limit theorem for bernoulli weighted random fields under hannan's condition. *Stochastic Processes and their Application*, 126:1819–1838.

Kolmogorov, A. and Rozanov, J. (1960). On strong mixing conditions for stationary gaussian processes. *Theo. Prob. Appl.*, 5:204–208.

Kontorovich, L. (2006). Metric and mixing sufficient conditions for concentration of measure. arxiv.org/abs/2102.03594.

Kontorovich, L. and Ramanan, K. (2008). Concentration inequalities for dependent random variables via the martingale method. *The Annals of Probability*, 36(6):2126–2158.

Lai, T. (1987). Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, pages 1091–1114.

Lai, T. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, pages 4–22.

Langford, J., Li, J., and Zhang, T. (2009). Sparse online learning via truncated gradient. *Journal of Machine Learning Research*, 10:777–801.

Lattimore, T. and Szepesvari, C. (2020). *Bandit algorithms*. Cambridge University Press.

Ledeoux, M. (1997). On talagrand's deviation inequalities for product measures. *ESAIM: Probability and Statistics*, 1:63–87.

Ledoux, M. (2001). The concentration of measure phenomenon. *American Mathematical Society*, pages 2–21.

Lin, J. and Cevher, V. (2018). Optimal convergence for distributed learning with stochastic gradient methods and spectral regularization algorithms. *Arxiv*, pages 1–53.

Littlestone, N. and Warmuth, M. (1994). The weighted majority algorithm. *Information and Computation*, 108(2):212–261.

Locatelli, A., Gutzeit, M., and Carpentier, A. (2016). An optimal algorithm for the thresholding bandit problem. In *Proceeding of the 33rd International Conference on Machine Learning*, volume 48, pages 1690–1698.

Loring, W. T. (2011). *An Introduction to Manifolds*. Springer New-York. 410 pages.

Lozano, A., Kulkarni, S., and Schapire, R. (2005). Convergence and consistency of regularized boosting algorithms with stationary $\beta-$mixing observations. In *NIPS'05 Proceedings of the 18th International Conference on Neural Information Processing Systems*, pages 819–826.

Mallet, V., Stoltz, G., and Mauricette, B. (2009). Ozone ensemble forecast with machine learning algorithms. *Journal of Geophysical Research: Atmospheres*, 114(D5).

Marton, K. (2004). Measure concentration for Euclidean distance in the case of dependent random variables. *The Annals of Probability*, 32(3):2526–2544.

Maume-Deschamps, V. (2006). Exponential inequalities and functional estimations for weak dependent data; applications to dynamical systems. *Stoch. Dyn.*, 6(4):535–560.

Mc Leish, D. (1975). Invariance principles and mixing random variables. *Econometric Theory*, 4:165–178.

McDiarmid, C. (1989). On the method of bounded differences. *Surveys in Combinatorics*, 141(1):148–188.

McDonald, D., Shalizi, C., and Schervish, M. (2011). Estimating beta-mixing coefficients. In *JMLR Workshop Conf Proc.*, volume 15, pages 516–524.

McDonald, D., Shalizi, C., and Schervish, M. (2015). Estimating beta-mixing coefficients via histograms. *Electronic Journal of Statistics*, 9(2):2855–2883.

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

Merlevede, F., Peligrad, M., and Rio, E. (2009). Bernstein inequality and moderate deviations under strong mixing conditions. *High dimensional probability*, 5:273–292.

Micchelli, C. A. and Pontil, M. (2004). A function representation for learning in banach spaces. In Shawe-Taylor, J. and Singer, Y., editors, *International Conference on Computational Learning Theory 17 (COLT 2004)*, pages 255–269.

Mohri, M. and Rostamizadeh, A. (2008). Rademacher complexity bounds for non-i.i.d. processes. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 21 (NIPS 2008)*, pages 1097–1104.

Mohri, M. and Rostamizadeh, A. (2010). Stability bounds for stationary ϕ-mixing and β-mixing processes. *JMLR*, 11:661–686.

Nahapetian, B. and Petrosian, A. (1992). Martingale-difference gibbs random fields and central limit theorems. *Annales Academiae Scientiarum Fennicae*, 17(1):105–110.

Narcowich, F. and Ward, J. (2004). Scattered-data interpolation on $\mathbb{R}^n$: error estimates for radial basis and band-limited functions. *SIAM J. MATH. ANAL*, 36:284–300.

Narcowich, F., Ward, J., and Wendland, H. (2004). Sobolev bounds on functions with scattered zeros, with applications to radial basis function surface fitting. *Mathematics of Computation*, 74:743–763.

Neu, G., György, A., Szepesvari, C., and Antos, A. (2013). Online markov decision processes under bandit feedback. In *IEEE Transactions on Automatic Control*, volume 59, pages 676–691.

Novak, E., Ulrich, M., Wozniakowski, H., and Zhung, S. (2017). Reproducing kernels of sobolev spaces on $\mathbb{R}^d$ and applications to embedding constants and tractability. *Arxiv*.

Ortner, R., Ryabko, D., Auer, P., and Munos, R. (2014). Regret bounds for restless markov bandits. *Theoretical Computer Science*, pages 62–76.

Pagliana, N., Rudi, A., De Vito, E., and Rosasco, L. (2020). Interpolation and learning with scale-dependent kernels. *Arxiv*.

Peligrad, M. (1983). A note on two measures of dependence and mixing sequences. *Adv. Appl. Probab.*, 15:461–464.

Peligrad, M. and Utev, S. (1997). Central limit theorem for linear processes. *The Annals of Probability*, 25(1):443–456.

Peligrad, M., Utev, S., and W.B., W. (2006). A maximal l_p inequality for stationary sequences and its applications. *Proceedings of the American Mathematical Society*, 135(2):541–550.

Peligrad, M., Utev, S., and Wu, W. (2007). A maximal l_p-inequality for stationary sequences and its applications. *Proceedings of the American Mathematical Society*, 135(2):541–550.

Perchet, V. and Rigollet, P. (2013). The multi-armed bandit problem with covariates. *Annals of Statistics*, 41(2):693–721.

Pillaud-Vivien, L., Rudi, A., and Bach, F. (2018). Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In *NIPS*.

Pinelis, I. (1992). An approach to inequalities for the distributions of infinite-dimensional martingales. *Probability in Banach Spaces*, 8:138–134.

Pinelis, I. (1994). Optimum bounds for the distributions of martingales in banach spaces. *Ann. Prob.*, 22:1679–1706.

Pinelis, I. and Sakhanenko, A. (1986). Remark on inequalities for large deviation probabilities. *Theory Probab. Appl.*, 30(1):143–148.

Potapov, D. and Sukochev, F. (2014). Fréchet differentiability of $\mathcal{S}^p$ norms. *Advances in Mathematics*, 262:436–475.

Rakhlin, A. and Sridharan, K. (2014). Online nonparametric regression. *Journal of Machine Learning Research*, pages 1–27.

Rakhlin, A., Sridharan, K., and A.Tewari (2014). Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and related random fields*, 161:111–153.

Rakhlin, A., Sridharan, K., and Tewari, A. (2015). Online learning via sequential complexities. *Journal of Machine Learning Research*, pages 155–186.

Rio, E. (1996). Sur le théorème de berry-esseen pour les suites faiblement dépendantes. *Probab. Th. Rel. Fields*, 104:255–282.

Rio, E. (2000). *Théorie asymptotique des processus aléatoires faibalement dépendants*. Springer, Berlin.

Rio, E. (2013). Extensions of the hoeffding-azuma inequalities. *Electron. Commun. Probab.*, 18(54):1–6.

Robbins, H. (1952). Some aspects of the sequential design of the experiments. *Bulletin of the American mathematical Society*, 58(5):527–535.

Rosasco, L., Belkin, M., and De Vito, E. (2010). On learning with integral operators. *Journal of Machine Learning Research*, (2):905–934.

Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Natl. Acad. Sci USA*, 42:43–47.

Rosenblatt, M. (2000). *Gaussian and Non-gaussian linear time series and random fields*. Springer, New York.

Rudi, A., Camoriano, R., and Rosasco, L. (2015). Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems*, pages 1657–1665.

Samson, P. (2000). Concentration of measure inequalities for markov chains and ϕ-mixing processes. *Ann. Prob.*, 28(1):416–461.

Sang, H. and Xiao, Y. (2018). Exact moderate and large deviations for linear random fields. *Journal of Applied Probability*, 55(55):431–449.

Schaback, J. (2007). Kernel-based meshless methods. Based on the lecture notes.

Schmidt, M. (2019). Machine learning for marketing. Blog.

Shalev-Shwartz, S. (2007). *Online Learning: Theory, Algorithms and Applications*. PhD thesis, The Hebrew University of Jerusalem.

Shirayev (1996). *Probability, 2nd edition*. Springer Science +Business Media, LLC.

Slivkins, A. (2019). *Introduction to multi-armed bandits*, volume 12. Foundations and trends in machine learning.

Smola, A. and Schölkopf, B. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, Cambridge, MA.

Song, G. and Zhang, H. (2011). Reproducing kernel banach spaces with ℓ_1 norm ii: Error analysis for regularized least squares regression. *Neural Computing*, 23:2713–2729.

Sriperumbudur, B., Fukumizu, K., and Lanckriet, G. (2011). Learning in hilbert vs. banach spaces: A measure embedding viewpoint. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 24 (NIPS 2011)*, pages 1773–1781.

Stein, E. (1970). *Singular integrals and differentiability properties of functions*. Princeton University Press.

Steinwart, I. (2009). Two oracle inequalities for regularized boosting classifiers. *Stat. Interface*, 2:271–284.

Steinwart, I. and Christmann, A. (2008). *Support vector machines*. Springer, New-York, 1 edition.

Talagrand, M. (1995). Concentration of measure and isoperimetric inequalties in product spaces. *Publications Mathematiques de l'I.H.E.S.*, 81:73–205.

Tekin, C. and Liu, M. (2010). Online algorithms for the multi-armed bandit problem with markovian rewards. In *48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE.

Thompson, W. (1933). On the likelihood that one unknown probability exceeds another in the view of the evidence of two samples. *Biometrika*, 25(3-4):285–294.

Tsybakov, A. (2009). *Introduction to nonparametric estimation*. Springer.

van de Geer, S. (2002). On hoeffding's inequality for dependent random variables. In *Empirical Process Techniques for Dependent Data*, pages 161–170. Birkhäuser Boston.

Vidyasagar, M. (2003). *Learning and Generlization: with Applications to Neural Networks*. Springer, New York, 1 edition.

Vovk, V. (1998). Competitive online linear regression. *Proceedings of the 1997 conference on advances in neural information processing systems, 10*, pages 364–370.

Vovk, V. (2001). Competitive online statistics. *International statistical review*, 69:213–248.

Vovk, V. (2006a). Metric entropy in competitive online prediction. *Arxiv*.

Vovk, V. (2006b). On-line regression competitive with reproducing kernel hilbert spaces. In *International Conference of Theory and Application of Models of Computation*, volume 69, pages 452–463.

Vovk, V. (2007). Competing with wild prediction rules. *Machine Learning*, 69:193–212.

Wendlandt, H. (2005). *Scattered Data Approximation*. Cambridge University Press.

Whittle, P. (1988). Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*.

Wintenberger, O. (2010). Deviation inequalities for sums of weakly dependent time series. *Electronic Communications in Probability*, (15):489–503.

Wu, W. (2005). Nonlinear system theory: another look on dependence. In *Proc. Natl. Acad. Sci.*, volume 102(40), pages 14150–14154. National Academy of Sciences.

Yu, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1):94–116.

Yurinskyi, V. (1970). On an infinite-dimensional version of s.n. bernstein's inequalities. *Theory Probab. Appl.*, (15):108–109.

Yurinskyi, V. (1995). *Sums and Gaussian Vectors*. Springer, Berlin.

Zadorozhnyi, O., Blanchard, G., and Carpentier, A. (2019). Restless dependent bandits with fading memory. arxiv.org/abs/1906.10454.

Zadorozhnyi, O., Gaillard, P., Gerchinovitz, S., and Rudi, A. (2021). Online nonparametric regression with Sobolev kernels. arxiv.org/abs/102.03594.

Zhang, H., Xu, Y., and Zhang, J. (2009). Reproducing kernel banach spaces for machine learning. *Journal of Machine Learning Research*, 10:2741–2775.

Zhang, H. and Zhang, J. (2013). Vector-valued reproducing kernel banach spaces with applications to multi-task learning. *Journal of Complexity*, 29:195–215.

Zhang, T. (2002). On the dual formulation of regularized learning schemes with convex risks. *Machine Learning*, 46:91–129.

Zhang, T. (2005). Learning bounds for kernel regression using effective data dimensionality. *Neural Computation 17(9)*, pages 2077–2098.

Zhdanov, F. and Kalnishkan, Y. (2010). An identity for kernel ridge regression. In *International Conference on Algorithmic Learning Theory*, pages 405–419. Springer.

Zinkewich, M. (2003). Online convex programming and generalized infinitesmal gradient descent. In *In Proceedings of the 20th International conference on Machine Learning*, page 9.

Zou, B., Zhang, H., and Xu, Z. (2009). Learning from uniformly ergodic Markov Chains. *Journal of Complexity*, 25(2):188–200.

Eigenständigkeitserklärung:

Ich versichere, die Arbeit selbstständig verfasst zu haben sowie keine anderen Quellen und Hilfsmittel als die im Literaturverzeichnis angegebenen verwendet zu haben.

Diese Arbeit ist bisher an keiner anderen Hochschule eingereicht worden.

Oleksandr Zadorozhnyi