Generalization bounds for nonparametric regression with β–mixing samples

David Barrera, Emmanuel Gobet

Generalization bounds for nonparametric regression with $\beta-$mixing samples

David Barrera [*†] Emmanuel Gobet [‡]

Abstract

In this paper we present a series of results that permit to extend in a direct manner uniform deviation inequalities of the empirical process from the independent to the dependent case characterizing the additional error in terms of beta-mixing coefficients associated to the training sample. We then apply these results to some previously obtained inequalities for independent samples associated to the deviation of the least-squared error in nonparametric regression to derive corresponding generalization bounds for regression schemes in which the training sample may not be independent.

These results provide a framework to analyze the error associated to regression schemes whose training sample comes from a large class of $\beta-$mixing sequences, including geometrically ergodic Markov samples, using only the independent case. More generally, they permit a meaningful extension of the Vapnik-Chervonenkis and similar theories for independent training samples to this class of $\beta-$mixing samples.

[*]Corresponding author.

[†]Email: `juandavid.barreracano@epfl.ch`. CMAP, École Polytechnique, Route de Saclay, 91128 Palaiseau cedex, France and SDS, École Polytechnique Fédérale de Lausanne. EPFL SB MATH MA C2 647 (Bâtiment MA) Station 8 CH-1015 Lausanne, Switzerland. Supported in 2019 by the Chaire Marchés en Mutation, Fédération Française Bancaire and by the Institut Louis Bachelier.

[‡]Email: `emmanuel.gobet@polytechnique.edu`. CMAP, Ecole Polytechnique, Route de Saclay, 91128 Palaiseau cedex, France. The authors research is part of the Chair *Financial Risks* of the *Risk Foundation* and the *Finance for Energy Market Research Centre*. This research also benefited from the support of the Chair *Stress Test, RISK Management and Financial Steering*, led by the French École Polytechnique and its Foundation and sponsored by BNP Paribas.

Contents

1 Introduction and background

This paper is a continuation of [BG19], where we addressed the problem of studying the error associated to a least-squares regression scheme in the nonparametric, distribution–free setting assuming that the training sample is independent.

1.1 The problem

Let $n \in \mathbb{N} := \{1, 2, \dots\}$ be a natural number (the "sample size"), let the "training sample" of "explanatory inputs" X_k and "responses" Y_k

$$D_n := ((X_k, Y_k))_{k\in\{1,\dots,n\}}$$

be a (not necessarily i.i.d.) random sequence in $S \times \mathbb{R}$, where S is a Polish space, defined on the probability space $(\Omega, \mathcal{E}, \mathbb{P})$, and let $\mathcal{F}_n$ be a family of Borel-measurable functions $S \to \mathbb{R}$ (the "space of hypotheses"). For $k \in \{1, \dots, n\}$, denote by $\mathbb{P}_{X_k}$ [respectively

$\mathbb{P}_{(X_k,Y_k)}$] the law of X_k [respectively (X_k, Y_k)], assume that $Y_k \in L^2_{\mathbb{P}_{X_k}}$, and let $\Phi_k : S \to \mathbb{R}$ be a version of the conditional expectation of Y_k given X_k, thus

$$\Phi_k(X_k) = \mathbb{E}\left[Y_k | X_k\right], \ \ \mathbb{P} - a.s.$$

Given such $(n, D_n, \mathcal{F}_n)$, a natural candidate to a "simultaneous" estimator within $\mathcal{F}_n$ of the regression functions Φ_k is the empirical regression function $\widehat{\Phi}_n$ defined as a solution to the *least-squares regression problem*

$$\widehat{\Phi}_n \in \arg\min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{k=1}^{n} |f(X_k) - Y_k|^2. \tag{1.1}$$

Indeed, by the orthogonal decomposition

$$\mathbb{E}\left[|Y_k - f(X_k)|^2\right] = \mathbb{E}\left[|Y_k - \Phi_k(X_k)|^2\right] + \mathbb{E}\left[|f(X_k) - \Phi_k(X_k)|^2\right],$$

the solutions Φ_n^* to the problem

$$\Phi_n^* \in \arg\min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{k=1}^{n} \mathbb{E}\left[|f(X_k) - \Phi_k(X_k)|^2\right] \tag{1.2}$$

are the same as those to the problem

$$\Phi_n^* \in \arg\min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{k=1}^{n} \mathbb{E}\left[|f(X_k) - Y_k|^2\right], \tag{1.3}$$

from where it follows that (1.1) and (1.2) are approximately the same problem *provided that the deviations of the random variables inside the* arg min *in* (1.1) *from their expectations inside the* arg min *of* (1.3) *are* (in some appropriate sense) *"negligible" uniformly in* $\mathcal{F}_n$.

In this context, the purpose of [BG19][1] was roughly speaking to show that, *when* D_n *is a sequence of independent random variables*, such deviations can be properly controlled provided a control on the complexity of $\mathcal{F}_n$[2] and a uniform bound of the response variables Y_k, and to describe some of the consequences of these controls for the problem of (weak and strong) rates and consistency, including the case where the response sequence $(Y_k)_k$ is not bounded. The innovation in [BG19] with respect to the classical i.i.d. case is, therefore, in the non-stationarity of D_n.

[1] Where we assumed $S = \mathbb{R}^d$, which is nonetheless largely irrelevant for the arguments.

[2] As measured typically by uniform entropy estimates, see Definition 3.5.

In continuation with this, we aim here at deriving some bounds for the probability of uniform deviations like

$$\mathbb{P}\left(\sup_{(g_1,\dots,g_n)\in\mathcal{G}_{1,\dots,n}}\frac{1}{n}\sum_{j=1}^{n}\left(ag_j(X_j,Y_j)+b\int g_j(x,y)\mathbb{P}_{X_j,Y_j}(\mathrm{d}x\,\mathrm{d}y)\right)\geq t\right) \tag{1.4}$$

when the training data D_n is not necessarily stationary, *nor* independent, but satisfies some β-mixing properties (particularly those in Definitions 2.7 and 2.8). Here a, b, t are scalar, $\mathcal{G}_{1,\dots,n}$ is a family of vectors $(g_1, \dots, g_n)$ whose entries are measurable functions $S \times \mathbb{R} \to \mathbb{R}$, and the complexity of $\mathcal{G}_{1,\dots,n}$ is controlled in the same ways as in [BG19].

We will show here how to "lift" the deviation inequalities in [BG19] from the independent to the dependent case using decoupling techniques associated to the $\beta-$ mixing coefficients of the training sample, and we will generalize some of the consequences for weak consistency and bounds on weak errors obtained in [BG19] for independent training samples using these ideas. When interpreted in the Markovian setting, these results provide error rates and consistency theorems for least-squares regression schemes under important ergodicity conditions on D_n. See for instance [TT96], [JR02], [DFG09], and the references therein.

1.2 Motivation

Our study is motivated in particular by the following application. In [FGM17], the authors investigate the numerical computation of the mean of a function of a conditional expectation in a rare-event regime, which takes the form

$$\mathcal{I} := \mathbb{E}\left[f(\tilde{X}, \mathbb{E}\left[Y|\tilde{X}\right])|\tilde{X}\in E_0\right],$$

where $\tilde{X}$ and Y are random variables, and the event $E_0 \in \mathcal{E}$ is rare (i.e. $\mathbb{P}(\tilde{X} \in E_0)$ small). This problem is prominent in financial/actuarial risk management when, as often, one has to deal with future risk exposure (modelled by $\mathbb{E}\left[Y|\tilde{X}\right] =: \Phi(\tilde{X})$) in extreme configurations (described by the set E_0). The above can be rewritten as $\mathcal{I} = \mathbb{E}\left[f(X, \mathbb{E}\left[Y|X\right])\right]$ where X has the conditional distribution of $\tilde{X}$ given $\{\tilde{X} \in E_0\}$. The computational strategy developed in [FGM17] consists in sampling n times (X, Y), computing the empirical regression function $\hat{\Phi}_n(x) \approx \mathbb{E}\left[Y|X = x\right]$ with these data, and averaging out the results over the explanatory sample $X_1, \dots, X_n$. One specific issue is that, E_0 being rare, naive i.i.d. sampling of X (with acceptance-rejection on E_0) is quite inefficient and one has to resort to a MCMC technique. The new $X_1, \dots, X_n$ are thus not independent, nor

stationary, but they fulfill some good β-mixing properties to ensure the approximation with respect to the (target) distribution of X. The convergence analysis is developed in [FGM17] and a upper bound on the Mean Square empirical norm

$$\mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n}\left(\hat{\Phi}_n(X_j)-\Phi(X_j)\right)^2\right]$$

is derived.

Using the current results of this work, we will be able to extend the scope of validity of the error analysis in [FGM17] in two directions: first, allowing the functions class for computing $\hat{\Phi}$ to be more general (and not only a linear space as in [FGM17]), including neural networks for instance; second, estimating the out-of sample error (as opposed to the in-sample error – aka empirical error).

1.3 Contributions of this paper

The results in this paper contribute to the existing literature mainly in two directions,

1. *A systematic presentation of the "lifting" of uniform deviation inequalities via Berbee's lemma.* This occupies Section 2, whose main results are Theorems 2.11 and Proposition 2.14.

   While the main purpose of this part of the paper is to permit a smooth and clear transition from some of the results under independence treated in [BG19] to the corresponding generalizations to dependence with $\beta-$mixing errors (achieved in Section 3), we aimed to present the results in this section in a manner that makes clear how these ideas go far beyond in generality than the kind of applications for which they are aimed at here. In this sense, we hope that they might serve as a useful reference for other works in which deviation inequalities for nonindependent sequences are sought for, provided that their independent counterparts are known or clearly obtainable.

2. *Weak rates and consistency theorems for least-squares regression schemes with non-independent training samples.* This part, developed in Section 3, consists in an application of the results from Section 2 to some of the results and proofs in [BG19].

   The conclusions obtained (see for instance Theorems 3.6 and 3.10) allow us to see how some the estimates obtained in [BG19] for independent samples generalize to

estimates for dependent samples[3] via the results from Section 2. These estimates are meaningful for a class of training samples with a kind of "superlinear $\beta-$mixing rate" (see (3.17) and (3.18)), providing in particular non-parametric, distribution-free estimates for geometrically ergodic Markovian training samples.

1.4 Background literature

Concentration and deviation inequalities for nonindependent samples constitute a topic of considerable research, in particular due to the importance of the Markovian case at the level of applications.

We start by mentioning [RM10], which uses basically the same coupling ideas developed in the present paper[4] to extend some of the inequalities in [GKKW02] for i.i.d. samples to the stationary $\beta-$mixing case and to describe the respective consequences for estimates of weak errors of least–squares regression schemes, including some penalisations. Our results give estimates that cover in the nonstationary case the corresponding estimates in [RM10] with a very significant improvement on the constants involved. These gains come in part from the work developed in [BG19].

We also mention [Ada08] (see also references therein). This paper presents first a deviation estimate ([Ada08, Theorem 4]) for independent samples under the assumptions that the functions in the space of hypotheses are centered with respect to the marginal laws of the sample and satisfy some bounds in terms of Orlicz norms, and then develops similar estimates ([Ada08, Theorems 6 and 7]) for uniformly bounded Markov samples under a certain "minorization condition" ([Ada08, Section 3.1]). In contrast with our results, the estimates for independent samples in [Ada08] cover cases in which the family of hypotheses is not uniformly bounded. Our estimates, on the other side, do not require the centering of the hypotheses with respect to the marginal laws in the independent case, and give rates for any exponentially $\beta-$mixing sequence of samples even if it is not Markovian, covering in particular the geometrically ergodic Markov chains in [Ada08]. We point out also that our applications (mainly Theorem 3.6) give bounds which are upper estimates on the probability of *some* individual large deviation of the empirical processes parametrized by the family of hypotheses from its corresponding mean, whereas the uniform estimates in [Ada08] ([Ada08, Theorems 4 and 7]) are rather estimates on the probability of deviations of the supremum of these empirical process from its mean: we will refer to these as "tail

[3] It is important to emphasize that, by reasons of space, in this process of generalizing we did not exhaust all the results available in [BG19]. The arguments for those treated here indicate how to extend the ones left aside.

[4] Our developments were indeed considerably inspired by the argument in [RM10].

estimates" in the rest of this section.

In [KM17], a coupling argument similar to the one in the present paper is used to address the problem of generalisation bounds for unspecified loss functions of regression algorithms in term of Rademacher complexities and $\beta-$mixing coefficients associated to dependences in the training sample, in a setting whose generality is approximately the same as that in our Section 2. The argument in [KM17], which proceeds via McDiarmid's inequality (see footnote 6 below), has the advantage of simplicity and generality compared to ours, but the rate obtained (roughly speaking $1/\sqrt{n}$ where n is the sample size) is suboptimal for the (square) loss function considered in our paper (we obtain roughly the rate $\log n/n$ in our analysis). For further comparison, notice again that our analysis does *not* proceed via tail estimates (see the comparison with [Ada08] before), and that we also cover the case of hypotheses depending on the index of the sample (the "time").

At a more ergodic theoretical level, let us mention the result in [DN93], where it is proved that the uniform convergence of averages holds for $\beta-$mixing samples (with stationary marginals) provided that it holds for i.i.d. samples with the same marginals when the class of functions in consideration has finite Vapnik-Chervonenkis (VC) dimension (as defined here in Example 3.7). Our paper can in part be considered a continuation of this story towards the investigation of rates of convergence, with more freedom in the independence assumption but with restrictions on the speed of mixing.

Let us comment briefly on the related research about these rates. Rates of uniform convergence to zero for the centered averages were for instance investigated in [Yuk86] (see also references therein), where the sample sequence is a $\phi-$mixing (and therefore $\beta-$mixing) process whose $\phi-$mixing coefficients satisfy certain growth conditions, and where the class of hypotheses is assumed to satisfy some "weak metric entropy" conditions and some controls on the associated maximal variance (see [Yuk86, Conditions (1.1)–(1.4), (1.6), and (1.8)–(1.10)]). Another instance of this story, closer to our paper, is [Yu94], which works under a general framework and via techniques that are quite similar to the ones here. It considers a case in which the sample sequence is $\beta-$mixing under a decay of the $\beta-$mixing coefficients that can be slower than ours, and it is also an interesting source of additional references. The results in [Yu94] complement our results in so far as [Yu94] considers slower mixing rates, and are complemented by our results in so far as [Yu94] relies on the assumption of stationary samples and time–independent spaces of hypotheses, which we dispense with here.

Like our own, many of the aforementioned papers proceed via comparisons with the corresponding results for the independent case and clever bounds on the additional error induced by dependence. The argument for the independent case typically depends on

estimates of probabilities like (1.4) when

$$\mathcal{G}_{1,\dots,n} = \{(g_1, \dots, g_n)\} \tag{1.5}$$

consists of a single point ("atomic estimates") and the training sequence D_n is independent, from where the uniform estimates (for more general $\mathcal{G}_{1,\dots,n}$) follow via finitely many applications of the atomic estimates using, for instance, "symmetrisation", "chaining", and estimates of covering or bracketing numbers ("entropy estimates"). See for instance [Pol90] for an introduction to these ideas.

These estimates have nonetheless been studied "directly" under classical dependence conditions in several works. The arguments in [Yuk86], for instance, depend on a result ([Yuk86, Lemma 2.1]) which is an extension to the $\phi-$mixing case of Bernstein's inequality.

But the developments in this directions have continued until recent years. One example is [MPR09] (see also references therein), whose results ([MPR09, Theorems 1 and 2]) imply that, if each g_j in (1.5) is bounded and $a = -b = 1$ (centered case), and if the $\alpha-$mixing coefficients associated to the sample sequence decay exponentially ([MPR09, Condition (1.3)][5]), then a Bernstein–type inequality bound holds (under (1.5)) at the right–hand side of (1.4). A second and final one is [DG15], where it is shown that, in the context of irreducible and aperiodic Markov chains, the assumption of geometric ergodicity is equivalent to the satisfaction of McDiarmid-type inequalities for separately bounded functionals of the observables[6] ([DG15, Theorem 2 and Remark 4]). One of the conclusions in [DG15] is that, for the small set specified in [DG15, Definition 1], these inequalities hold (also) under the conditional law at every starting point in such set and for the deviations of the expectation with respect to such conditional law.

For the case of suprema of partial sums, the results explained in Section 2.6 are comparable with those in [DG15]: they give analogous consequences for the probability of large

[5] This condition is weaker than (2.13) below for $\gamma = 1$, but we remind that the estimates in [MPR09] are not uniform.

[6] If $(X_k)_k$ is the Markov chain in consideration, this amounts to the satisfaction of estimates of the type

$$\mathbb{P}\left(|K(X_1, \dots, X_n) - \mathbb{E}\left[K(X_1, \dots, X_n)\right]| > t\right) \leq C_1 \exp\left(-C_2 t^2 / \sum_{k=1}^{n} L_k^2\right)$$

where $K : \mathbb{R}^n \to \mathbb{R}$ is any (Borel-measurable) function such that $x \mapsto K(x_1, \dots, x_{k-1}, x, x_{k+1}, \dots, x_n)$ is bounded by $L_k > 0$ when $x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n$ is fixed, for every $k \in 1, \dots, n$. Notice in particular that this covers tail estimates like those in [Ada08] and [KM17] when the entries of $\mathcal{G}_{1,\dots,n}$ are uniformly bounded. For potential comparisons of [DG15] with our results see again the comparison with [Ada08] and [KM17] above.

deviations[7] which rely only on the rate of decay of the $\beta-$mixing coefficients associated to the underlying sequence and on the corresponding estimates from the independent case. These estimates admit therefore as a special case that in which the training sample comes from a Markov chain as those in [DG15].

Organization of the paper. The rest of the paper is organized as follows: we begin Section 2 by introducing some notational conventions that will be used in the forthcoming pages. We explain next, also in Section 2, how to transport uniform deviation inequalities from the independent to the dependent case estimating the additional error via the $\beta-$mixing coefficients. Section 3 presents some applications to problems in nonparametric least–squares regression under dependent training samples, in continuity with some of the independent-case considerations in [BG19].

2 Bridge between $\beta-$mixing and independent sequences

Our strategy for deriving concentration-of-measure inequalities for dependent sequences is to leverage on decoupling techniques and deviation inequalities for independent sequences (as those of [BG19]). These inequalities with dependent sequences will take the form of Lemma 2.1 and Proposition 2.14, which constitute the main result of this section. The derivation is made in several steps.

2.1 Notation and conventions

The following conventions will be used in this paper

- We depart from a probability space $(\Omega, \mathcal{E}, \mathbb{P})$ supporting all the random variables that will appear in our statements and proofs (the existence of this space can be verified *a posteriori*).
- We denote by $\mathbb{N} = \{1, 2, \dots, \}$ the set of positive integers.
- For $k, n \in \mathbb{N}$, we will sometimes denote $k : n := \{k, \dots, n\}$ ($k : n := \emptyset$ if $k > n$), and we use the notation $c_{1:n}$ for a sequence ($n-$tuple) of elements $(c_1, \dots, c_n)$.
- More generally, given a subset $J \subset \mathbb{N}$, $c_J := (c_j)_{j \in J}$ denotes a sequence indexed by J, which we will call a $J-$*tuple*. The cardinality of J is denoted by $|J|$. If $c_J = (c_j)_{j \in J}$

[7] As opposed, again, to the tail estimates that follow from [DG15].

is given and $J' \subset J$, we will denote *the projection of* c_J *onto the* J' *coordinates* by $c_{J'}$[8]. Thus for $c_J := (c_j)_{j\in J}$,

$$c_{J'} = (c_j)_{j\in J'}. \tag{2.1}$$

- For a subset $J \subset \mathbb{N}$ and a family of sets $\{C_j\}_{j\in J}$ indexed by J, we use the notation

$$C_J^{\otimes} := \{c_J = (c_j)_{j\in J} | \, \forall j \in J : c_j \in C_j\} \tag{2.2}$$

for *the product of the* C_j*'s*[9].

- Sometimes[10] we will deal with *sets* $\mathcal{F}_J$ *of* J*-tuples* which are not necessarily a product of sets. In *all* of these cases the indexing set (i.e., J) of the elements of $\mathcal{F}_J$ will be indicated in the notation. In analogy with (2.1), given such $\mathcal{F}_J$ and $J' \subset J$, $\mathcal{F}_{J'}$ denotes the projection of $\mathcal{F}_J$ into the J' coordinates

$$\mathcal{F}_{J'} := \{f_{J'} : f_J \in \mathcal{F}_J\}, \tag{2.3}$$

where each $f_{J'}$ is given by (2.1). Thus for instance, for the set in (2.2) and $J' \subset J$, we have $\left(C_J^{\otimes}\right)_{J'} = C_{J'}^{\otimes}$.

- We reserve the character S for Polish spaces with variations from taking products as in the above, and we will usually denote by Z a generic random vector in S with compatible variations when S is a product space. Thus Z_J typically denotes a random element of a product space $S_J^{\otimes}$. This is, $Z_J = (Z_j)_{j\in J}$ with $Z_j : \Omega \to S_j$ $\mathcal{E}-$measurable and S_j a Polish space.

- If Z is a random element of S and B a Borel set of S, we use the standard notation $\{Z \in B\} := \{\omega \in \Omega : Z(\omega) \in B\}$ for the preimage of B (which is a set in $\mathcal{E}$). We use similarly the standard notation $\mathbb{P}_Z$ for the law of Z: given a Borel set $B \subset S$,

$$\mathbb{P}_Z(B) := \mathbb{P}(\{Z \in B\}).$$

- For a Polish space S, $\mathcal{L}_S$ denotes the space of Borel-measurable functions $S \to \mathbb{R}$. If $\{S_j\}_{j\in J}$ ($J \subset \mathbb{N}$) are Polish spaces, set $\mathcal{L}_S^{\otimes_J} := \Pi_{j\in J}\mathcal{L}_{S_j}$. A subset of $\mathcal{L}_S^{\otimes_J}$ will

[8] Of course, we will be careful to use properly the notation to avoid confusions: in no place we will for instance denote two different tuples as c_J and $c_{J'}$, except if their entries with index in $J \cap J'$ are equal.

[9] The same care will be taken to avoid confusion here: we will always use the same character (here "C.") for the sets involved in the product.

[10] Especially for function hypotheses, see for instance (2.4) below.

be called a *sequential family of functions compatible with* $S_J^{\otimes}$, or simply a *sequential family of functions* when there is no ambiguity for $\{S_j\}_{j\in J}$. One relevant example is the sequential family of functions

$$\mathcal{G}_{\mathcal{F},1:n} := \{g_{f,1:n} : f \in \mathcal{F}\} \tag{2.4}$$

in (3.7).

- When needed, we will operate with sequential families of functios in a componentwise manner, thus given $f_J = (f_j)_j, f'_J = (f'_j)_j$ in $\mathcal{F}_J$, where $\mathcal{F}_J$ is a sequential family of functions, $f_J + f_{J'} = (f_j + f'_j)_j$, $f_J f'_J := (f_j f'_j)_j$, $|f_J| := (|f_j|)_j$, and so on.

- A couple $(Z_J, \mathcal{F}_J)$ where Z_J is a random element of $S_J^{\otimes}$ and $\mathcal{F}_J$ is a sequential family of functions compatible with $S_J^{\otimes}$ is called a *composable pair*. In this definition, the reference to $S_J^{\otimes}$ is implicit and omitted for the sake of convenience. Notice that if $(Z_J, \mathcal{F}_J)$ is a composable pair and $J' \subset J$, then $(Z_{J'}, \mathcal{F}_{J'})$ is a composable pair.

- The *empirical mean* and the *average mean* associated to the composable pair (Z_J, f_J), denoted respectively by $A_{Z_J} f_J$ and $\mu_{Z_J} f_J$ are defined, for nonempty finite J, as

$$A_{Z_J} f_J := \frac{1}{|J|} \sum_{j\in J} f_j(Z_j), \qquad \mu_{Z_J} f_J = \frac{1}{|J|} \sum_{j\in J} \int_{S_j} f_j(z)\, \mathbb{P}_{Z_j}(\mathrm{d}z).$$

  (the second average is defined only for those f_J where it makes sense, including the possible value ∞). With this convention, we will use the short notation

$$(aA_{Z_J} + b\mu_{Z_J}) f_J := aA_{Z_J} f_J + b\mu_{Z_J} f_J = \frac{a}{|J|} \sum_{j\in J} f_j(Z_j) + \frac{b}{|J|} \sum_{j\in J} \int_{S_j} f_j(z)\, \mathbb{P}_{Z_j}(\mathrm{d}z),$$

  for any real constants a, b.

- When convenient, we identify a function f with the *constant* sequence of functions $(f_j)_{j\in J}$ where $f_j = f$ for all $j \in J$, which together with the above permits, for instance, an unambiguous interpretation of the object "$\mu_{Z_J} f$".

2.2 "Union bound" for deviations of averages

We begin with the following elementary lemma, which shows that estimates on the distribution function associated to suprema of (generally non–centered) empirical means can be obtained from corresponding estimates on the empirical means over the indexes in a partition of the set $\{1, \ldots, n\}$.

Lemma 2.1 ("Union bound" for deviation of averages)**.** *Let $n \in \mathbb{N}$, let $\mathcal{J}$ be a partition (by nonempty subsets) of $\{1,\dots,n\}$ and let $(Z_{1:n}, \mathcal{G}_{1:n})$ be a composable pair. Then for every $(a,b,t) \in \mathbb{R}^3$*

$$\mathbb{P}\left(\sup_{g_{1:n}\in\mathcal{G}_{1:n}} (aA_{Z_{1:n}} + b\mu_{Z_{1:n}})g_{1:n} \geq t\right) \leq \sum_{J\in\mathcal{J}} \mathbb{P}\left(\sup_{g_J\in\mathcal{G}_J} (aA_{Z_J} + b\mu_{Z_J})g_J \geq t\right). \quad (2.5)$$

Proof. The proof is easy: for every $J \in \mathcal{J}$, denote $\gamma_J := |J|/n$. Notice that $\sum_{J\in\mathcal{J}} \gamma_J = 1$. With this, (2.5) is an immediate consequence of the subadditivity of the supremum, linearity, and the union bound:

$$\left\{\sup_{g_{1:n}\in\mathcal{G}_{1:n}} (aA_{Z_{1:n}} + b\mu_{Z_{1:n}})g_{1:n} \geq t\right\} = \left\{\sup_{g_{1:n}\in\mathcal{G}_{1:n}} \sum_{J\in\mathcal{J}} \gamma_J(aA_{Z_J} + b\mu_{Z_J})g_J \geq \sum_{J\in\mathcal{J}} \gamma_J t\right\}$$
$$\subset \left\{\sum_{J\in\mathcal{J}} \gamma_J \sup_{g_J\in\mathcal{G}_J} (aA_{Z_J} + b\mu_{Z_J})g_J \geq \sum_{J\in\mathcal{J}} \gamma_J t\right\} \subset \bigcup_{J\in\mathcal{J}} \left\{\sup_{g_J\in\mathcal{G}_J} (aA_{Z_J} + b\mu_{Z_J})g_J \geq t\right\}.$$

□

The above lemma shows that if we can find appropriate subsampling partition $\mathcal{J}$ for which we have an exponential (for instance) inequality for the deviation probability, the same type of inequality holds for the full sample $\{1,\dots,n\}$. The construction of the partition $\mathcal{J}$ will be made using the β-mixing properties of the sequence $Z_{1:n}$, which is now discussed.

Remark 2.2 (Generalization under a convex-like estimate)**.** Lemma 2.1 can clearly be extended to any family of (Borel-measurable) functionals $\{K_J\}_{J\subset\mathbb{N}}$, $K_J : S_J^{\otimes} \to \mathbb{R}$ with the property that for every disjoint family $\{J_1,\dots,J_r\} \subset 2^{\mathbb{N}}$ and some nonnegative $\gamma_1,\dots,\gamma_r$ with $\sum_k \gamma_k = 1$, $K_J(Z_J) \leq \sum_k \gamma_k K_{J_k}(Z_{J_k})$, $\mathbb{P}-$a.s., where $J := \cup_k J_k$. For such a family one has the inequality

$$\mathbb{P}\left(K_J(Z_J) \geq t\right) \leq \sum_{k=1}^{r} \mathbb{P}\left(K_{J_k}(Z_{J_k}) \geq t\right),$$

for every $t \in \mathbb{R}$, every $J \subset \mathbb{N}$, and every partition $J_1,\dots,J_r$ of J. See also Remark 2.12 below.

2.3 The $\beta-$mixing coefficients

In this section, we introduce some facts about $\beta-$mixing coefficients that will be useful later. For an account on mixing properties, we refer the reader to [DDL+07, Dou12, DMPS19].

2.3.1 Basic definitions and properties

Definition 2.3 ($\beta-$mixing coefficients)**.** *Let $\mathcal{E}_1$ and $\mathcal{E}_2$ be two sub-sigma algebras of $\mathcal{E}$. The $\beta-$mixing coefficient $\beta(\mathcal{E}_1, \mathcal{E}_2)$ between $\mathcal{E}_1$ and $\mathcal{E}_2$ is defined as*

$$\beta(\mathcal{E}_1, \mathcal{E}_2) := \mathbb{E}\left[\operatorname*{ess\,sup}_{E_1 \in \mathcal{E}_1} |\mathbb{P}(E_1) - \mathbb{P}[E_1|\mathcal{E}_2]|\right]. \tag{2.6}$$

For a definition of the essential supremum, "ess sup", of a family of random variables, see [Nev75, Proposition VI-1-1]. It follows in particular that there exists a countable family $\{E_{1,n}\}_n \subset \mathcal{E}_1$ such that

$$\beta(\mathcal{E}_1, \mathcal{E}_2) = \mathbb{E}\left[\sup_n |\mathbb{P}(E_{1,n}) - \mathbb{P}[E_{1,n}|\mathcal{E}_2]|\right]. \tag{2.7}$$

Remark 2.4 (A characterization. Properties.)**.** If $\{E_{1,n}\}_n$ is the family in (2.7) and $\mathcal{E}_2$ is countably generated, then

$$\begin{aligned}\beta(\mathcal{E}_1, \mathcal{E}_2) &= \mathbb{E}\left[\sup_n |\mathbb{P}E_{1,n} - \mathbb{P}\left(E_{1,n}|\mathcal{E}_2\right)|\right] \\ &= \frac{1}{2} \sup_{(P_1,P_2)\in\mathcal{P}_{\mathcal{E}_1}\times\mathcal{P}_{\mathcal{E}_2}} \sum_{(E_1',E_2')\in P_1\times P_2} |\mathbb{P}\left(E_1'\right)\mathbb{P}\left(E_2'\right) - \mathbb{P}\left(E_1' \cap E_2'\right)|, \end{aligned} \tag{2.8}$$

where $\mathcal{P}_{\mathcal{E}_k}$ $(k = 1, 2)$ denotes the family of finite partitions of Ω by $\mathcal{E}_k-$sets[11]. This representation holds in particular if $\mathcal{E}_k := \sigma(Z_k)$ is the sigma algebra generated by Z_k, where Z_k $(k = 1, 2)$ is a random element of a Polish space S_k.

Additionally, it follows that

(i) *The $\beta-$mixing coefficients are symmetric: $\beta(\mathcal{E}_1, \mathcal{E}_2) = \beta(\mathcal{E}_2, \mathcal{E}_1)$.*

(ii) *$\beta(\cdot, \cdot)$ is increasing in each component: if $\mathcal{E}_k' \subset \mathcal{E}_k$ $(k = 1, 2)$ then*

$$\beta(\mathcal{E}_1', \mathcal{E}_2') \leq \beta(\mathcal{E}_1, \mathcal{E}_2). \tag{2.9}$$

[11] This can be seen for instance by noticing that there exist increasing families of finite fields $\{\mathcal{E}_{j,k}\}_k$ $(j = 1, 2)$ with $\cup_k \mathcal{E}_{j,k} \subset \mathcal{E}_j$ such that

$$\beta(\mathcal{E}_1, \mathcal{E}_2) = \lim_k \lim_l \beta(\mathcal{E}_{1,l}, \mathcal{E}_{2,k}),$$

and using elementary considerations on $\beta(\mathcal{E}_1, \mathcal{E}_2)$ when $\mathcal{E}_j$ are finite fields. For a proof under slightly more restrictive hypotheses, see [DMPS18, Proposition F.2.8].

(iii) $\beta(\mathcal{E}_1, \mathcal{E}_2) = 0$ *if and only in* $\mathcal{E}_1$ *and* $\mathcal{E}_2$ *are* $\mathbb{P}-$*independent.*

The first two properties follow by the equality between the extreme sides of (2.8). The third one is clear even from the general definition (2.6).

2.3.2 $\beta-$ coefficients of $m-$dependence

We now extend the previous considerations to a case involving families of sub-sigma algebras related to a sequence of random variables $Z_{1:\infty}$. The aim is to set a precise discussion involving some β-mixing coefficients associated to "the present" and "the past" of this sequence.

Definition 2.5 ($\beta-$ coefficients of $m-$dependence)**.** *Given a subset* $J \subset \mathbb{N}$*, a random element* Z_J *of* $S_J^{\otimes}$*, and* $(m, l) \in \mathbb{N} \times \mathbb{N}$*, the* $l-$*th* $\beta-$*coefficient of* $m-$*dependence of* Z_J *is defined as*

$$\beta_{Z_J}(m, l) := \beta(\sigma(Z_{J\cap[1,l-m]}), \sigma(Z_{J\cap\{l\}})),$$

where the right–hand side is defined in (2.6) *and with the convention* $Z_\emptyset := \emptyset$*. The maximal* $\beta-$*coefficient of* $m-$*dependence is denoted by* $\beta_{Z_J}(m)$*:*

$$\beta_{Z_J}(m) := \sup_{l\in\mathbb{N}} \beta_{Z_J}(m, l). \tag{2.10}$$

Thus for $l \in J$, $\beta_{Z_J}(m, l)$ gives the $\beta-$mixing coefficient between Z_l and the "distant past" (at least m units before l) of Z_J. Similarly, $\beta_{Z_J}(m)$ is the smallest upper bound of the $\beta-$mixing coefficients of Z_J within "some present" and its (at least m units) "distant past".

We list, for future reference, some properties of $\beta_{Z_\cdot}(\cdot, \cdot)$ and $\beta_{Z_\cdot}(\cdot)$.

Properties 2.6 (of β_{Z_J})**.**

1. *If* $l \notin J$ *then* $\beta_{Z_J}(m, l) = 0$ *for all* m*, see* (iii) *in Remark 2.4.*
2. *For fixed* l*,* $\beta_{Z_J}(\cdot, l)$ *is decreasing. Thus* $(\beta_{Z_J}(m))_m$ *is also decreasing.*
3. *A sufficient condition for* $\beta_{Z_J}(m) = 0$ *is the* $m-$*dependence of* Z_J*, i.e., the hypothesis that for every* l*,* $Z_{J\cap[1,l]}$ *and* $Z_{J\cap[l+m,\infty)}$ *are independent (this condition is not necessary*[12]*). In particular,* $\beta_{Z_J}(\cdot) \equiv 0$ *if the entries of* Z_J *are independent.*

[12] Choose random variables X, Y, Z with X independent of Y and X independent of Z but with $Y + Z$ not independent of X, choose X' independent of $\sigma(X, Y, Z)$ and consider, for $n = 4$, $J = \{1, 2, 3, 4\}$, and $m = 2$, the choices $Z_1 = X$, $Z_2 = X'$, $Z_3 = Y$, $Z_4 = Z$).

4. *If $J' \subset J$, then $\beta_{Z_{J'}}(\cdot,\cdot) \leq \beta_{Z_J}(\cdot,\cdot)$ (pointwise) by (2.9). If in particular $Z_{1:n}$ is a random element of $S_{1:n}^{\otimes}$ and $J \subset \{1,\dots,n\}$ is any subset then*

$$\beta_{Z_J}(\cdot,\cdot) \leq \beta_{Z_{1:n}}(\cdot,\cdot). \tag{2.11}$$

5. *Assume that the partition $\mathcal{J}$ of $\{1,\dots,n\}$ is such that the indexes within each $J \in \mathcal{J}$ are separated by a "minimal gap", say $1 \leq m < n$,*[13] *then the inequality*

$$\beta_{Z_J}(m') \leq \beta_{Z_J}(1) \leq \beta_{Z_{1:n}}(m) \tag{2.12}$$

holds for all $m' \in \mathbb{N}$ and all $J \in \mathcal{J}$. The first inequality follows from Property 2, the second follows from Property 1, the inequality (2.11), and the fact that for all $J \in \mathcal{J}$ and all $l \in J$, $J \cap [1, l-1] \subset \{1,\dots,l-m\}$.

2.3.3 Examples

Of particular interest for us are the following mixing hypotheses on the rate of decay of $\beta_{Z_\cdot}(\cdot)$.

Definition 2.7 (Sub exponentially $\beta-$mixing process)**.** *Let $Z_{1:\infty}$ be a random element of $S_{1:\infty}^{\otimes}$. $Z_{1:\infty}$ is subexponentially $\beta-$mixing with parameters $(a,b,\gamma) \in (0,\infty)\times(0,\infty)\times(0,\infty)$ if for all $m \in \mathbb{N}$*

$$\beta_{Z_{1:\infty}}(m) \leq a\exp(-bm^{\gamma}). \tag{2.13}$$

Definition 2.8 (Subpolinomially $\beta-$mixing processes)**.** *Let $Z_{1:\infty}$ be a random element of $S_{1:\infty}^{\otimes}$. $Z_{1:\infty}$ is subpolinomially $\beta-$mixing with parameters $(a,\gamma) \in (0,\infty)\times(1,\infty)$ if for all $m \in \mathbb{N}$*

$$\beta_{Z_{1:\infty}}(m) \leq am^{-\gamma}. \tag{2.14}$$

If $Z_{1:\infty}$ is a Markov Chain with state space S, then, by the Markov property,

$$\beta_{Z_{1:\infty}}(m) = \sup_n \beta(\sigma(Z_n), \sigma(Z_{n+m})).$$

Sufficient conditions for exponentially mixing rates (i.e. (2.13) holds with $\gamma \geq 1$) of Markov chains can be consulted also in [Bra05], [FM03a], [MT09, Chapter 16]. For sufficient conditions implying (in the Markovian setting) subexponential $\beta-$mixing rates (2.13) with $\gamma \in (0,1)$ or polynomial rates like (2.14), see for instance [TT96], [JR02], [FM03b], [DFMS04], [DFG09], and the references therein.

[13] I.e., m is the smallest m' such that, for any $J \in \mathcal{J}$ and any different $j_1, j_2 \in J$, $|j_1 - j_2| \geq m'$.

2.4 Berbee's Lemma

The $\beta-$mixing coefficients measure, on a certain sense, the "($\mathbb{P}-$)distance from independence" between two sigma–algebras. This notion is put forward in a more concrete way by the following classical coupling result[14] (see [Ber79, Corollary 4.2.5], [Dou12, Theorem 1, p.7]).

Lemma 2.9 (Berbee's Lemma)**.** *Let (V, W) be a random vector in $S_1 \times S_2$. There exists a S_2-valued random vector W^*, distributed as W, independent of V, and with the property*

$$\beta(\sigma(V), \sigma(W)) = \mathbb{P}(W \neq W^*).$$

The above lemma admits the following (apparently) generalised version[15]:

Lemma 2.10 (Generalised Berbee's Lemma)**.** *Given $N \in \mathbb{N}$ and a random sequence $V_{1:N}$ of $S_{1:N}^{\otimes}$, there exists a random sequence $V_{1:N}^*$ with independent entries such that for every $1 \leq k \leq N$,*

1. *V_k and V_k^* have the same distribution, and*
2.

$$\mathbb{P}(V_k \neq V_k^*) = \beta(\sigma(V_{1:k-1}), \sigma(V_k)). \tag{2.15}$$

(In particular, $V_1 = V_1^*$, $\mathbb{P}-$a.s.)

Proof. We start with a preliminary observation: notice that if V_1, V_2, V are random variables with V independent of $\sigma(V_1, V_2)$ then for any Borel set $E_2 \subset S_2$,

$$\mathbb{P}\left(V_2 \in E_2 | \sigma(V_1)\right) = \mathbb{P}\left(V_2 \in E_2 | \sigma(V_1, V)\right),$$

$\mathbb{P}-$a.s. Using the characterization in the first equality of (2.8) (with obvious adjustments on notation) and the symmetry of $\beta(\cdot, \cdot)$,

$$\beta(\sigma(V_1, V), \sigma(V_2)) = \mathbb{E}\left[\sup_n \left|\mathbb{P}\left(V_2 \in E_{2,n}\right) - \mathbb{P}\left(V_2 \in E_{2,n} | \sigma(V_1, V)\right)\right|\right]$$

[14] We omit specifications about the "richness" of $(\Omega, \mathcal{A})$, which are implicitly embedded in the introductory remarks.

[15] Whose proof, although developed independently, follows an argument resembling the one in [Vie97, p.484].

$$=\mathbb{E}\left[\sup_n \left|\mathbb{P}\left(V_2 \in E_{2,n}\right) - \mathbb{P}\left(V_2 \in E_{2,n} | \sigma(V_1)\right)\right|\right] \qquad = \beta(\sigma(V_1), \sigma(V_2)). \tag{2.16}$$

Now we prove the statement. First, we assume that $N \geq 2$ (otherwise the conclusion is trivial, even without the vacuous property of independence, for $V_1^* := V_1$).

Let now $V_{1:N}$ be a random sequence in $S_{1:N}^{\otimes}$. We will construct a sequence $V_{2:N}^*$ satisfying, for all $1 \leq k < N$, the property **P(k)** defined by

P(k): The sequence $V_{k+1:N}^*$ is such that, for $k \leq j < N$,

1. V_{j+1} and V_{j+1}^* are identically distributed with $\beta(\sigma(V_{1:j}), \sigma(V_{j+1})) = \mathbb{P}(V_{j+1} \neq V_{j+1}^*)$.
2. The vectors $V_{1:j}$ and $V_{j+1:N}^*$ are independent.

which is easily seen to be sufficient to prove the claim of Lemma 2.10 by defining $V_1^* := V_1$.

We will construct $V_{2:N}^*$ by backward induction: start defining V_N^* by applying Lemma 2.9 with $V = V_{1:N-1}$ and $W = V_N$. This verifies the satisfaction of **P(N-1)**.

Now, assume that **P(k)** has been verified by $V_{k+1:N}^*$ for some $1 \leq k < N$. An application of Berbee's lemma with $V := (V_{1:k-1}, V_{k+1:N}^*)$ and $W := V_k$ guarantees the existence of a random variable V_k^* distributed as V_k and independent of $\sigma(V_{1:k-1}, V_{k+1:N}^*)$ such that

$$\mathbb{P}(V_k \neq V_k^*) = \beta(\sigma(V_{1:k-1}, V_{k+1:N}^*), \sigma(V_k)) = \beta(\sigma(V_{1:k-1}), \sigma(V_k)),$$

where the last equality follows by an application of (2.16). The augmented sequence $V_{k:N}^*$ satisfies therefore **P(k-1).** After $N-1$ steps this gives the desired construction. □

2.5 A general estimate for decoupled averages

Our next result, Theorem 2.11, is an inequality relating the distribution function of certain random variables defined by suprema and associated to a composable pair $(Z_{1:n}, \mathcal{G}_{1:n})$ to the corresponding distribution functions over sets of indexes in a partition $\mathcal{J}$ of $\{1, \ldots, n\}$ and a "decoupling" of $Z_{1:n}$ over each one of the set of indexes in $\mathcal{J}$.

Indeed, the inequality (2.5), combined with Lemma 2.10, allows us to relate distribution functions as in the left–hand side of (2.5) to a sum of similar distribution functions *which are defined for independent sequences*, controlling the additional error with the $\beta-$dependence coefficients associated to $\mathcal{J}$ in Definition 2.5.

Theorem 2.11. *Let $\mathcal{J}$ and $(Z_{1:n}, \mathcal{G}_{1:n})$ be as in Lemma 2.1. There exists a sequence $Z_{1:n}^*$ with the following properties:*

1. *For every $k \in \{1, \dots, n\}$, the distributions of Z_k^* and Z_k are the same.*

2. *For every $J \in \mathcal{J}$, Z_J^* is an independent sequence.*

3. *The inequality*

$$\begin{aligned} &\mathbb{P}\left(\sup_{g_{1:n} \in \mathcal{G}_{1:n}} (aA_{Z_{1:n}} + b\mu_{Z_{1:n}}) g_{1:n} \geq t \right) \\ &\qquad \leq \sum_{J \in \mathcal{J}} \left(\mathbb{P}\left(\sup_{g_J \in \mathcal{G}_J} (aA_{Z_J^*} + b\mu_{Z_J}) g_J \geq t \right) + \sum_{k \in J} \beta_{Z_J}(1,k) \right). \end{aligned} \tag{2.17}$$

holds for every $(a, b, t) \in \mathbb{R}^3$.

Proof. We start by an application of Lemma 2.1, the next step is a further estimate of the right–hand side of (2.5) via Lemma 2.10.

Indeed, fix $J \in \mathcal{J}$. We apply Lemma 2.10 to Z_J to construct Z_J^*: if $J := \{j_1, \dots, j_N\}$ in increasing order, construct Z_J^* by replacing $V_k := Z_{j_k}$ in Lemma 2.10.

Properties 1. and 2. are immediate from this construction. Notice also that, by the construction and (2.15),

$$\mathbb{P}(Z_k \neq Z_k^*) = \beta_{Z_J}(1,k), \qquad \forall k \in J. \tag{2.18}$$

Now, using the inclusion

$$\left\{ \sup_{g_J \in \mathcal{G}_J} (aA_{Z_J} + b\mu_{Z_J}) g_J \geq t \right\} \subset \left\{ \sup_{g_J \in \mathcal{G}_J} (aA_{Z_J^*} + b\mu_{Z_J}) g_J \geq t \right\} \cup \bigcup_{k \in J} \{Z_k \neq Z_k^*\},$$

(2.17) follows from the union bound via Lemma 2.1 and (2.18). □

Remark 2.12 (Complement to Remark 2.2)**.** Let $\{K_J\}_J$ be a family of functionals as in Remark 2.2, then exactly the same argument as in the proof of Theorem 2.11 gives that, for every finite $I \subset \mathbb{N}$ and every partition $\mathcal{J}$ of I

$$\mathbb{P}\left(K_I(Z_I) \geq t\right) \leq \sum_{J \in \mathcal{J}} \left(\mathbb{P}\left(K_J(Z_J^*) \geq t\right) + \sum_{k \in J} \beta_{Z_J}(1,k) \right), \tag{2.19}$$

where Z_I^* satisfies properties 1. and 2. above (with $\{1, \dots, n\}$ replaced by I).

Remark 2.13 (Relationship with Bernstein's method)**.** Let $\{K_J\}_J$ be again as in Remark 2.2. Given a finite set $J \subset \mathbb{N}$, a random element Z_J of $S_J^{\otimes}$, a partition $\{J_1, \ldots, J_r\}$ of J, and a partition $I_1, \ldots, I_s$ of $\{1, \ldots, r\}$, denote, for every $k \in \{1, \ldots, r\}$, $J(I_k) := \cup_{s \in I_k} J_s$. Then using an argument similar to the one in the proof of Theorem 2.11 it is easy to prove that

$$\mathbb{P}\left(K_J(Z_J) \geq t\right) \leq \sum_{k=1}^{r} \left(\mathbb{P}\left(K_{J(I_k)}(Z^{**}_{J(I_k)}) \geq t\right) + \sum_{j \in I_k} \beta_{(Z_{J_l})_{l \in I_k}}(1, j) \right) \tag{2.20}$$

where

1. Z_{J_k} and $Z^{**}_{J_k}$ have the same distribution, for $k \in 1, \ldots, r$.
2. For every fixed $k \in \{1, \ldots, r\}$, the sequence $(Z^{**}_{J_l})_{l \in I_k}$ is independent.

Assume that $n = 2am$ for some $(a, m) \in \mathbb{N} \times \mathbb{N}$. Then one can group the sequence $Z_{1:n}$ into $2m$ disjoint blocks $J_1, \ldots, J_{2m}$ of successive elements, each of length a, and classify the blocks in "odd" and "even" blocks, which corresponds in (2.20) to taking I_1 and I_2 as (respectively) the odd and even numbers in $\{1, \ldots, 2m\}$. An application of (2.20) (for $J = \{1, \ldots, n\}$) together with an easy adaptation of (2.12) gives the estimate (with a slight abuse of notation)

$$\mathbb{P}\left(K(Z_{1:n}) \geq t\right) \leq \sum_{k=1}^{2} \mathbb{P}\left(K((Z^{**}_{J_l})_{l \in I_k}) \geq t\right) + 2m\beta_{Z_{1:n}}(a),$$

where, for fixed k, each sequence $(Z^{**}_{J_l})_{l \in I_k}$ is independent. This is the key idea in "Bernstein's partition method" (see for instance [KM17] and the references therein).

The most important difference between the partitions used in the estimates (2.19) and (2.20) is that, in (2.19), there is dependence *within* the blocks $(Z^*_J)_{J \in \mathcal{J}}$ but there is independence *inside* each block Z^*_J. In (2.20) the situation is somewhat reversed: *for fixed* $k \in \{1, \ldots, r\}$, there is independence *within* the blocks $(Z^{**}_{J_l})_{l \in I_k}$, but there is dependence *inside* each block $Z^{**}_{J_l}$. The reader is invited to consider the consequences of this difference for what follows.

2.6 Abstract lifting of deviation inequalities

In this concluding part we present a result indicating how to "lift" deviation inequalities from the independent to the (possibly) dependent case via Theorem 2.11. The purpose of this result for what follows is to serve as an intermediate step towards the beta–mixing

generalization of the deviation estimates proved in [BG19]. The notation and conventions are those explained in Section 2.1.

Proposition 2.14 (Abstract lifting of deviation inqualities)**.** *Let $n \in \mathbb{N}$, let $(a, b, B) \in \mathbb{R} \times \mathbb{R} \times (0, \infty]$, and let $(Z_{1:n}, \mathcal{G}_{1:n})$ be a composable pair such that*

$$\sup_{g_{1:n} \in \mathcal{G}_{1:n}} \sup_{1 \le k \le n} ||g_k(Z_k)||_{\mathbb{P},\infty} \le B.$$

Moreover, assume that there exists a function

$$L_{a,b} : \{1, \dots, n\} \times [0, \infty] \to [0, \infty)$$

such that for any $t \ge 0$, $J \subset \{1, \dots, n\}$ and some Z_J^* with independent entries and the same marginals as Z_J, we have

$$\mathbb{P}\left(\sup_{g_J \in \mathcal{G}_J} (aA_{Z_J^*} + b\mu_{Z_J}) g_J \ge t \right) \le L_{a,b}(|J|, t). \tag{2.21}$$

Let $m \in \{1, \dots, n\}$ and write

$$n := qm + r, \quad \text{with } q = \left\lfloor \frac{n}{m} \right\rfloor, \quad 0 \le r < m$$

for the Euclidean algorithm for n divided by m; then the estimate

$$\begin{aligned} &\mathbb{P}\left(\sup_{g_{1:n} \in \mathcal{G}_{1:n}} (aA_{Z_{1:n}} + b\mu_{Z_{1:n}}) g_{1:n} \ge t \right) \\ &\quad \le (rL_{a,b}(q+1, t) + (m-r)L_{a,b}(q, t) + n\beta_{Z_{1:n}}(m)) \mathbf{1}_{\{t \le (|a|+|b|)B\}} \\ &\quad \le (m(L_{a,b}(q+1, t) \vee L_{a,b}(q, t)) + n\beta_{Z_{1:n}}(m)) \mathbf{1}_{\{t \le (|a|+|b|)B\}} \end{aligned} \tag{2.22}$$

holds (with the convention $L_{a,b}(n+1, t) \equiv L_{a,b}(n, t)$).

The inequality (2.21) as an assumption is quite standard: it says that for a given class of functions $\mathcal{G}_J$, the uniform deviation depends of the size of the sample $|J|$ and the amplitude of the deviation t, consistently with many results on uniform deviation inequalities (see for instance [GKKW02, LT13]). As we will see later, the complexity of the class $\mathcal{G}_J$ typically appears in $L_{a,b}(.)$.

Proof. The second inequality in (2.22) is trivial. We proceed to prove the first inequality. In view of the Euclidean decomposition $n = qm + r$, consider the *$m-$steps partition*

$$\mathcal{J}_{m\,steps} := \{J_1, \dots, J_m\}$$

of $\{1,\dots,n\}$ specified by

$$J_k := \big\{k + lm\big\}_{l=0}^{q}, \qquad 1 \le k \le r$$
$$J_k := \big\{k + lm\big\}_{l=0}^{q-1}, \qquad r < k \le m.$$

In words, J_k is the set obtained by starting from k and moving to the right in steps of m units as far as possible before quitting the set $\{1,\dots,n\}$. Clearly, $|J_k| = q+1$ for $1 \le k \le r$ and $|J_k| = q$ for $r < k \le m$.

We apply Theorem 2.11 with $\mathcal{J} := \mathcal{J}_{m\,steps}$. This gives the upper bound

$$\left(rL_{a,b}(q+1,t) + (m-r)L_{a,b}(q,t) + \sum_{k=1}^{m}\sum_{j\in J_k} \beta_{Z_{J_k}}(1,j) \right) \mathbf{1}_{\{t\le(|a|+|b|)B\}}$$

for the left–hand side of (2.22). The conclusion follows using the estimate (2.12) which gives $\sup_{j\in J_k}\beta_{Z_{J_k}}(1,j) \le \beta_{Z_{J_k}}(1) \le \beta_{Z_{1:n}}(m)$ and the fact that $\sum_{k=1}^{m}\sum_{j\in J_k} 1 = n$. □

3 Some applications to nonparametric regression

In this section, we develop some of the applications of the results in Section 2 to the problems addressed, in the context of independent samples, within [BG19]. The notation, again, comes from Section 2.1.

3.1 Empirical covering numbers

The functions $L_{a,b}$ in (2.21) usually depend on the complexity of the functions class $\mathcal{G}_J$, through its covering number w.r.t. a suitable semimetric, see the seminal work [VC71]. We now recall the notion of *r-coverings* and *covering numbers*, taking care of extending it to our case of sequences of spaces $\mathcal{G}_J$.

Definition 3.1 ($r-$covering, covering numbers)**.** *Let $(\mathcal{G}, d)$ be a semimetric space, let $\mathcal{G}_0 \subset \mathcal{G}$, and let $r \in [0,\infty)$. An $r-$covering of $\mathcal{G}_0$ with respect to d is a set $\mathcal{G}' \subset \mathcal{G}$ with the property that, for every $g \in \mathcal{G}_0$, there exists $g' \in \mathcal{G}'$ satisfying*

$$d(g,g') < r.$$

The $r-$covering number of $\mathcal{G}_0$ with respect to d is defined as

$$\mathcal{N}^{(d)}(r,\mathcal{G}_0) := \min\{|\mathcal{G}'| : \ \mathcal{G}' \subset \mathcal{G} \text{ is an } r-\text{covering of } \mathcal{G}_0 \text{ with respect to } d\}.$$

Notice that the meaning of $\mathcal{N}^{(d)}(r, \mathcal{G}_0)$ depends not only on the set $\mathcal{G}_0$ and the metric $d|_{\mathcal{G}_0 \times \mathcal{G}_0}$, but also on the space $\mathcal{G}$ where d is defined.

The following type of covering numbers are of special relevance for us.

Definition 3.2 (Empirical covering numbers)**.** *Let $J \subset \mathbb{N}$ be a finite set, let $\mathcal{G}_J \subset \mathcal{L}_S^{\otimes_J}$ be a sequential family of functions, and let $z_J \in S_J^{\otimes}$ be given. We define the empirical L_1 $r-$covering numbers of $\mathcal{G}_J$ at z_J, $\mathcal{N}_1(r, \mathcal{G}_J, z_J)$, as*

$$\mathcal{N}_1(r, \mathcal{G}_J, z_J) := \mathcal{N}^{(d^1_{z_J})}(r, \mathcal{G}_J),$$

where $d^1_{z_J}$ is the empirical L^1-seminorm $d^1_{z_J}(g_J, g'_J) := A_{z_J}|g_J - g'_J|$ on the product space $\mathcal{L}_S^{\otimes_J}$.

Remark 3.3 (Measurability issues)**.** It is clear that, if Z_J is a random element of $S_J^{\otimes}$, $\omega \mapsto \mathcal{N}_1(r, \mathcal{G}_J, Z_J(\omega))$ is a nonnegative function. To avoid unnecessary measurability discussions, we will denote by $\mathbb{E}[\mathcal{N}_1(r, \mathcal{G}_J, Z_J)]$ the *outer* expectation of $\mathcal{N}_1(r, \mathcal{G}_J, Z_J)$:

$$\mathbb{E}[\mathcal{N}_1(r, \mathcal{G}_J, Z_J)] := \inf_h \mathbb{E}[h],$$

where the infimum is taken over the random variables $h : \Omega \to \mathbb{R}$ with $\mathcal{N}_1(r, \mathcal{G}_J, Z_J) \leq h$ (except on a set of $\mathbb{P}-$measure zero), with the convention $\inf \emptyset = \infty$.

3.2 Uniform deviation inequalities for dependent samples

We start by recalling the following result, which is a consequence of [BG19, Theorem 2.2] (with easy simplifications left to the reader). We will use it as a "toy" theorem, whose extension to the dependent case will illustrate some arguments that are not written in detail later.

Theorem 3.4 (Uniform deviation probability, independent version)**.** *Let $X_{1:n}$ be a random element of $(\mathbb{R}^d)^n$ with independent entries, and assume that $(X_{1:n}, \mathcal{F}_{1:n})$ is a composable pair where $\mathcal{F}_{1:n}$ is a pointwise measurable sequential family*[16] *with $f_k : \mathbb{R}^d \to [0, B]$ $(k = 1, \dots, n)$ for some $B > 0$ and for each $f_{1:n} \in \mathcal{F}_{1:n}$. Then for*

$$(\varepsilon, c, \gamma, \gamma') \in \times(0,1) \times (1, \infty) \times (1, \infty) \times (1, \infty),$$

the estimate

$$\mathbb{P}\left(\sup_{f_{1:n} \in \mathcal{F}_{1:n}} ((1-\varepsilon)A_{X_{1:n}} - (1+\varepsilon)\mu_{X_{1:n}}) f_{1:n} > t \right)$$

[16] I.e. such that there exists $\{f^{(k)}_{1:n}\}_k \subset \mathcal{F}_{1:n}$ with the property that, for every $f_{1:n} \in \mathcal{F}_{1:n}$, there exists a sequence $(k_l)_l$ satisfying $\lim_l f^{(k_l)}_{1:n} = f_{1:n}$ pointwise.

$$\leq \frac{2\gamma}{\gamma-1}\mathbb{E}\left[\mathcal{N}_1(\frac{1}{2}u_1(c,\gamma')\,t,\mathcal{F}_{1:n},X_{1:n})\right]\exp(-\frac{1}{2B}u_2(c,\gamma')\varepsilon nt)$$

holds with

$$u_1(c,\gamma') := (1-\frac{1}{c})\frac{1}{\gamma'}, \qquad u_2(c,\gamma') := (1-\frac{1}{c})^2(1-\frac{1}{\gamma'}), \tag{3.1}$$

provided that

$$t \geq \frac{Bc}{2}\left(\frac{\gamma}{n}\right)^{1/2}.$$

Our extension of this result will be made with the help of the following notion, which we will discuss briefly in Section 3.3:

Definition 3.5 (Uniform L^1-entropy estimates)**.** *Let $J \subset \mathbb{N}$ and let $\mathcal{G}_J \subset \mathcal{L}_S^{\otimes_J}$ be given. A Borel-measurable function $\lambda : \mathbb{N}\times(0,\infty)\to[1,\infty]$ is called an empirical L_1-uniform entropy estimate of $\mathcal{G}_J$ (or simply, a uniform entropy estimate of $\mathcal{G}_J$) if for every finite subset $J'\subset J$ and every $r\in(0,\infty)$*

$$\log\left(\sup_{z_{J'}\in S_{J'}^{\otimes}}\mathcal{N}_1(r,\mathcal{G}_{J'},z_{J'})\right)\leq\lambda(|J'|,r).$$

Going back to the extension of Theorem 3.4, assume the existence of a uniform entropy estimate λ of $\mathcal{G}_{1:n}$, then λ is clearly a uniform entropy estimate of $\mathcal{G}_J$ for every $J\subset\{1,\ldots,n\}$; in addition, for any random element Z_J of $S_J^{\otimes}$, we have

$$\mathbb{E}\left[\mathcal{N}_1(r,\mathcal{G}_J,Z_J))\right]\leq\exp(\lambda(|J|,r)).$$

Consequently, under the hypotheses of Theorem 3.4, we have that the inequality

$$\begin{aligned}&\mathbb{P}\left(\sup_{f_J\in\mathcal{F}_J}((1-\varepsilon)A_{X_J}-(1+\varepsilon)\mu_{X_J})f_J>t\right)\\&\quad\leq\frac{2\gamma}{\gamma-1}\exp\left(-\frac{1}{2B}u_2(c,\gamma')\varepsilon|J|t+\lambda\Big(|J|,\frac{1}{2}u_1(c,\gamma')\,t\Big)\right)=:L_{c,\gamma,\gamma',\varepsilon}(|J|,t)\end{aligned}\tag{3.2}$$

holds for every $J\subset\{1,\ldots,n\}$, provided this time that $t\geq\frac{Bc}{2}(\frac{\gamma}{|J|})^{1/2}$. We can "hide" this restriction on t by extending (3.2) to the estimate

$$\mathbb{P}\left(\sup_{f_J\in\mathcal{F}_J}((1-\varepsilon)A_{X_J}-(1+\varepsilon)\mu_{X_J})f_J>t\right)\leq\mathbf{1}_{\{t<\frac{Bc}{2}\left(\frac{\gamma}{|J|}\right)^{1/2}\}}+L_{c,\gamma,\gamma',\varepsilon}(|J|,t)\mathbf{1}_{\{t\geq\frac{Bc}{2}\left(\frac{\gamma}{|J|}\right)^{1/2}\}},$$

which holds for every $J\subset\{1,\ldots,n\}$ and every $t>0$, always under the hypotheses of Theorem 3.4. This, together with Proposition 2.14, allows us to deduce the following "$\beta-$version" of Theorem 3.4.

Theorem 3.6 (Uniform deviation probability, $\beta-$version)**.** *Let $X_{1:\infty}$ be a random sequence in $(\mathbb{R}^d)^{\mathbb{N}}$. For $n \in \mathbb{N}$, assume that $(X_{1:n}, \mathcal{F}_{1:n})$ is a composable pair where $\mathcal{F}_{1:n}$ is a pointwise measurable sequential family and each $f_{1:n} \in \mathcal{F}_{1:n}$ is a sequence of functions with $f_k : \mathbb{R}^d \to [0, B]$ $(k = 1, \ldots, n)$ for some $B > 0$. Assume that λ is a uniform entropy estimate of $\mathcal{F}_{1:n}$ (Definition 3.5), and let*

$$(\varepsilon, c, \gamma, \gamma') \in (0, 1) \times (1, \infty) \times (1, \infty) \times (1, \infty).$$

Then, with u_j $(j = 1, 2)$ as in (3.1), with $L_{c,\gamma,\gamma',\varepsilon} : \{1, \ldots, n\} \times [0, \infty) \to [0, \infty)$ as in (3.2) and with $\beta_{X_{1:\infty}}(\cdot)$ as in (2.10), the estimate

$$\begin{aligned}
&\mathbb{P}\left(\sup_{f_{1:n} \in \mathcal{F}_{1:n}} ((1-\varepsilon)A_{X_{1:n}} - (1+\varepsilon)\mu_{X_{1:n}}) f_{1:n} \geq t \right) \\
&\quad \leq \left[m\Big(L_{c,\gamma,\gamma',\varepsilon}\Big(\Big\lfloor \frac{n}{m} \Big\rfloor, t \Big) \vee L_{c,\gamma,\gamma',\varepsilon}\Big(\Big\lfloor \frac{n}{m} \Big\rfloor + 1, t \Big) \Big) + n\beta_{X_{1:\infty}}(m) \right] \mathbf{1}_{\{t \leq 2B\}},
\end{aligned}$$

holds for every $m \in \{1, \ldots, n\}$ (with the convention $L_{c,\gamma,\gamma',\varepsilon}(n+1, t) \equiv L_{c,\gamma,\gamma',\varepsilon}(n, t)$), provided that

$$t \geq \frac{Bc}{2} \left(\frac{\gamma}{\lfloor \frac{n}{m} \rfloor} \right)^{1/2}.$$

In practice, the choice of m will depend on the applications at hand. Typically, it will be done with the goal of minimizing in a convenient way the deviation from the rates obtained in the independent case (towards optimal extensions of the results in [BG19]) that follow from the results under consideration (see for instance Propositions 3.16 and 3.17 below). The uniform deviations proved in [BG19] for independent samples ([BG19, Section 2]) can be extended in a similar manner.

3.3 Remarks on entropy estimates

The definition of uniform entropy estimates, given here as a uniform estimate of the covering numbers associated to the L^1- empirical seminorm (Definition 3.2), can of course be extended via Definition 3.1 to other families of semimetrics in $\mathcal{L}_S^{\otimes_J}$, such as the empirical L^p-seminorms $(p \geq 1)$ defined as $d^p_{z_J} := (A_{z_J}|f_J - g_J|^p)^{1/p}$. The relationships between covering numbers for different semimetrics can be relevant: for instance, it is clear from the Cauchy-Schwarz inequality that $d^1_{z_J}(\cdot, \cdot) \leq d^2_{z_J}(\cdot, \cdot)$, which implies an analogous inequality for the respective covering numbers of the same sequential family $\mathcal{G}_J \subset \mathcal{L}_S^{\otimes_J}$.

It is also important for applications to describe the stability of covering numbers with respect to some elementary operations between families of functions, see for instance [GKKW02, Lemmas 6.3, 6.4 and 6.5], [Pol90, Section 5], and [vW96, Theorem 2.6.9]. These translate to analogous "stability properties" for uniform entropy estimates.

Let us now give some instances of the notion of entropy estimates.

Example 3.7 (VC dimension. The "Sauer-Shelah" estimate.)**.** One important instance of uniform entropy estimates, which is part of the framework used in [BG19], is the following: for a function $f : \mathbb{R}^d \to \mathbb{R}$, define the *subgraph of* f as the set

$$G_f^+ := \{(x, y) \in \mathbb{R}^d \times \mathbb{R} : y \leq f(x)\}.$$

The *VC-dimension* $V_{\mathcal{F}}$ *of a family* $\mathcal{F}$ *of functions* $\mathbb{R}^d \to \mathbb{R}$ is the supremum of the natural numbers l with the following property: there exists a set $G \subset \mathbb{R}^d \times \mathbb{R}$ with l elements such that every subset $G' \subset G$ can be written in the form $G' = G \cap G_f^+$ for some $f \in \mathcal{F}$.

When $\mathcal{F}$ is a family of bounded, nonnegative functions $f : \mathbb{R}^d \to [0, B]$, one has the following uniform L^1-entropy estimate ([GKKW02, Lemma 9.2 and Theorem 9.4.]) for the "diagonal" family

$$\mathcal{F}_{1:n} := \{\underbrace{(f, \dots, f)}_{n \text{ times}} : f \in \mathcal{F}\} \tag{3.3}$$

of hypotheses on $\mathcal{F}$, which we typically identify with $\mathcal{F}$ itself:[17] for $r \in [0, B/4]$, every $J \subset \{1, \dots, n\}$, and every $z_J \in (\mathbb{R}^d)^J$,

$$\begin{aligned} \log(\mathcal{N}_1(r, \mathcal{F}_J, z_J)) &\leq \lambda_{V_{\mathcal{F}},B}(r) \\ &:= \log 3 + V_{\mathcal{F}}(1 + \log 2 + \log(B/r) + \log(1 + \log 3 + \log(B/r))), \end{aligned} \tag{3.4}$$

which is clearly $O(\log(1/r))$ as $r \to 0^+$ when $V_{\mathcal{F}} < \infty$[18], in particular when $\mathcal{F} = T_B\mathcal{H}$ is the family of truncated functions (see Section 3.4.1) from a vector space of dimention $d_{\mathcal{H}} < \infty$, thanks to the bounds

$$V_{T_B\mathcal{H}} \leq V_{\mathcal{H}} \leq d_{\mathcal{H}} + 1$$

([GKKW02, Theorem 9.5 (and previous paragraph) and Equation (10.23)]).

[17] Covering numbers for non-diagonal families are nonetheless implicit within what follows, for instance in the arguments behind (3.13).

[18] Note also that the restriction $r \in [0, B/4]$ can be easily bypassed: one can for instance take $\lambda_{V_{\mathcal{F}},B}(r) = 0$ if $r > B$, and for $B \in [B/4, B]$, one can take $\lambda_{V_{\mathcal{F}},B}(r) := \lambda_{V_{\mathcal{F}},4B}(r)$, where $\lambda_{V_{\mathcal{F}},4B}(r)$ is defined as in (3.4) (valid for $r \in [0, 4B/4] = [0, B]$). A similar trick allows us to give uniform entropy estimates via (3.4) on (perhaps nonpositive) families $\mathcal{F}$ of functions $f : \mathbb{R}^d \to [-B, B]$: the family $\mathcal{F}' = \mathcal{F}+B := \{f+B : f \in \mathcal{F}\}$ has the same covering numbers as $\mathcal{F}$, satisfies $V_{\mathcal{F}} = V_{\mathcal{F}'}$, and its elements are functions $f : \mathbb{R}^d \to [0, 2B]$.

The estimate (3.4) is a consequence of the celebrated *Sauer-Shelah lemma* ([Sau72],[She72]). It is therefore a relationship between the *complexity of* $\mathcal{F}$, as measured by $V_{\mathcal{F}}$, and the notion of uniform entropy estimates.

There are other notions of complexity for families of functions, also associated to uniform entropy estimates, that are very relevant within the current literature, such as the (distribution–dependent) *Rademacher complexity* and the *fat shattering dimension*. See [MR08] and [RST15] for respective discussions beyond the i.i.d. case.

Example 3.8 (Neural networks)**.** A second example is given by neural networks: it is shown in [GKKW02, p.314] that if $\sigma : \mathbb{R} \to [0,1]$ is any cumulative distribution function (for instance a "sigmoid" function with asymptotes $y = 0$ and $y = 1$) and $\mathcal{F}$ is the family of functions $f : \mathbb{R}^d \to \mathbb{R}$ of the form

$$f(x) = b_0 + \sum_{k=1}^{N} b_k \sigma(u_k^T x + a_k)$$

with $N \in \mathbb{N}$ fixed, and with $((a_k)_k, (b_k)_k, (u_k)_k) \in \mathbb{R}^N \times \mathbb{R}^{N+1} \times (\mathbb{R}^d)^N$ subject to the restriction $\sum_k |b_k| \leq B$ for some $B > 0$, then the corresponding diagonal family $\mathcal{F}_{1:n}$ (see (3.3)) satisfies

$$\log(\mathcal{N}_1(r, \mathcal{F}_J, z_J)) \leq ((2d+5)N + 1)(1 + \log(12) + \log(B/r) + \log(N+1)) \quad (3.5)$$

for every $r \in (0, B/2)$.

Notice that the estimates in Examples 3.7 and 3.8 do not depend on $|J|$. In our applications, the dependence on $|J|$ will be introduced by lower-bounding the radius $r \geq r(|J|)$ where these estimates are applied.

Remark 3.9 (Additional comments on $V_{\mathcal{F}} < \infty$)**.** Restricting the analysis to the case $V_{\mathcal{F}} < \infty$ is basically a convenience due the estimate (3.4) for the quantitative bounds on the errors discussed in our applications, but one can extend these to some cases of "infinite complexity" ($V_{\mathcal{F}} = \infty$) using similar estimates.

One instance is the estimate (3.5) for neural networks (see [Son92] for examples showing that $V_{\mathcal{F}}$ can be infinite within this context), but as indicated in [BG19, Remarks 3.3, 3.5 and 3.19], one can extend the applications below to cases in which the left–hand side of (3.4) is bounded by a function of the form $O((1/r)^\alpha)$ $(r \to 0^+)$ for some $\alpha \in (0,1)$.

3.4 Weak least-squares error estimates under dependence

In what follows, we provide some applications of the results above to distribution-free and nonparametric error bounds associated to schemes based in the method of least-squares regression.

3.4.1 Least-squares setting

We recover the following definitions and conventions from [BG19]:

- *Truncation Operator.* First, we remind the *truncation operator,* defined for a constant $B > 0$ and associating to any real–valued function g the function $T_B g$ defined as

$$T_B g(x) = \max\{\min\{g(x), B\}, -B\}.$$

- *Least–squares regression (LSR) objects.* Consider a random vector $(X, Y)_{1:\infty}$ of $(\mathbb{R}^d \times \mathbb{R})^{\mathbb{N}}$, assume that for all k, $Y_k \in L^2_{\mathbb{P}}$, and pick a version $\Phi_k : \mathbb{R}^d \to \mathbb{R}$ of $\mathbb{E}[Y_k|X_k]$. Thus

$$\Phi_k(X_k) = \mathbb{E}[Y_k|X_k], \quad \mathbb{P} - a.s., \qquad k = 1, \dots, n.$$

  For a fixed $n \in \mathbb{N}$, consider the data $D_n := (X, Y)_{1:n}$. Given a family $\mathcal{F}$ of Borel-measurable functions $\mathbb{R}^d \to \mathbb{R}$, let $\widehat{\Phi}_n = \widehat{\Phi}_n(\mathcal{F}, D_n)$ be a solution (assume it exists) of the least–squares regression problem associated to $\mathcal{F}$ and D_n: if we identlfy

$$f \equiv \underbrace{(f, \dots, f)}_{n \text{ times}} \tag{3.6}$$

  for $f \in \mathcal{F}$ (compare with (3.3)), then

$$\widehat{\Phi}_n \in \arg\min_{f \in \mathcal{F}} A_{(X,Y)_{1:n}} |f - y_{1:n}|^2,$$

  where $y_{1:n} = (y, \dots, y)$ with $y : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ the projection on the second coordinate ($y(x_0, y_0) = y_0$), and where we naturally identify $f \in \mathcal{F}$ with the function $\mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ whose value at (x, y) is $f(x)$.

  We also write $\Phi_{1:n} := (\Phi_k)_{k=1}^n$. Notice that, consistently with the (3.6)

$$\widehat{\Phi}_n \equiv \underbrace{(\widehat{\Phi}_n, \dots, \widehat{\Phi}_n)}_{n \text{ times}}.$$

- *Pointwise deviations of the least–squares error.* In this context, we reserve a special notation for the family $\mathcal{G}_{\mathcal{F},1:n} = \{g_{f,1:n}\}_{f \in \mathcal{F}}$ whose elements are the sequential functions

$$g_{f,1:n} := |y_{1:n} - f|^2 - |y_{1:n} - \Phi_{1:n}|^2. \tag{3.7}$$

  From here, the meaning of $\mathcal{G}_{\mathcal{F},J}$ for any $J \subset \{1, \dots, n\}$ is clear (see (2.3)).

3.4.2 A weak L^2-error estimate for dependent samples

We continue with the following "$\beta-$version" of [BG19, Theorem 3.1]. The setting is that in Section 3.4.1:

Theorem 3.10. ($\beta-$version of [BG19, Theorem 3.1]). *Assume that $\mathcal{F}$ is a pointwise measurable class of functions with associated $VC-$dimension $V_{\mathcal{F}} < \infty$, and that $||Y_k||_{\mathbb{P},\infty} \leq B$ for some $B > 0$ and all k. Assume further that $(c, \lambda, n, m) \in (1, \infty) \times (1, \infty) \times \mathbb{N} \times \mathbb{N}$ are such that*

$$\lambda \leq \frac{3 + \sqrt{1 + 8c}}{4}, \qquad \left\lfloor \frac{n}{m} \right\rfloor \geq \exp\left(\frac{c^2 - 71}{4V_{\mathcal{F}}} \right), \tag{3.8}$$

(in particular $n \geq m$), then the estimate

$$\begin{aligned} \mathbb{E}\left[\mu_{X_{1:n}} | T_B \widehat{\Phi}_n - \Phi_{1:n} |^2 \right] \leq & \frac{B^2}{\left\lfloor \frac{n}{m} \right\rfloor} \theta_0 \left(1 + \theta_1 + V_{\mathcal{F}}(\theta_2 + \log(\theta_2)) \right) \\ & + 16B^2(1 + \lambda) n \beta_{(X,Y)_{1:\infty}}(m) + \lambda \inf_{f \in \mathcal{F}} \mu_{X_{1:n}} |f - \Phi_{1:n}|^2 \\ & (= \underbrace{\text{“Variance” + “}\beta-\text{mixing error”}}_{\text{“Statistical error”}} + \text{“scaled bias”}.) \end{aligned} \tag{3.9}$$

holds, where

$$\begin{aligned} \theta_0 =& \theta_0(\lambda, c) := 32 \left(\frac{1}{3}(1 - \frac{1}{c})(1 - \frac{1}{\lambda}) + (2\lambda - 1) \right)^2 (\frac{c}{c-1})^3 \frac{\lambda}{\lambda - 1}, \\ \theta_1 =& \theta_1(c, m) := \log(6(c+1)(2c+3)) + \log m, \\ \theta_2 =& \theta_2(c, n, m) := 1 + \log 24 + \log(1 + \sqrt{1 + \frac{c(c+1)}{\left\lfloor \frac{n}{m} \right\rfloor + 1}}) - \log(c - \frac{1}{c}) + \log(\left\lfloor \frac{n}{m} \right\rfloor + 1). \end{aligned}$$

Remark 3.11 (A simplified version of the variance in (3.9))**.** It is easy to see that for every $c', \lambda' > 1$, there exists a constant $C_{c',\lambda'} > 0$ such that

$$\begin{aligned} \frac{1}{C_{c',\lambda'}} \frac{V_{\mathcal{F}}}{\lambda - 1}(1 + \log n - \log m) \leq & \theta_0 \left(1 + \theta_1 + V_{\mathcal{F}}(\theta_2 + \log(\theta_2)) \right) \\ \leq & C_{c',\lambda'} \frac{V_{\mathcal{F}}}{\lambda - 1}(\log c + \log n), \end{aligned} \tag{3.10}$$

provided that $(n, c, \lambda) \in \mathbb{N} \setminus \{1\} \times (c', \infty) \times (1, \lambda')$. We shall when convenient write the variance term in (3.9) as

$$O\left(\frac{B_n^2 V_{\mathcal{F}_n}(\log c_n + \log n)}{(\lambda_n - 1)n} m_n \right), \tag{3.11}$$

with the "right" of letting $n \to \infty$ as far as $\{\lambda_n\}_n \subset (1,\infty)$ is bounded and $\{c_n\}_n \subset (1,\infty)$ is away from 1, and provided that (3.8) holds for the parameters $(c_n, \lambda_n, V_{\mathcal{F}_n}, m_n)$.

Remark 3.12 (The "$\beta-$mixing error–variance" tradeoff)**.** Notice also that reducing m simultaneaously increases the $\beta-$mixing error and reduces the variance in (3.9). Our choice of m in the applications below is based on a qualitatively optimal tradeoff between these errors: the tradeoff is made for $m = m_n$ with the goal of minimizing the distance to the smallest possible statistical error, achieved in the independent case in which $\beta_{(X,Y)_{1:\infty}}(1) = 0$ and the statistical error is therefore equal to the variance term in (3.9) for $m = 1$.

Proof of Theorem 3.10. First, as proved in [BG19, Section 3.2], we have the estimate

$$\mathbb{E}\left[\mu_{X_{1:n}}|T_B\widehat{\Phi}_n - \Phi_{1:n}|^2\right] \leq \mathbb{E}\left[\left(\sup_{f\in T_B\mathcal{F}}(\mu_{(X,Y)_{1:n}} - \lambda A_{(X,Y)_{1:n}})g_{f,1:n}\right)^+\right] + \lambda \inf_{f\in\mathcal{F}} \mu_{X_{1:n}}|f - \Phi_{1:n}|^2 \tag{3.12}$$

We proceed now to bound conveniently the distribution function $[0,\infty) \to [0,1]$ defined by

$$t \mapsto \mathbb{P}\left(\sup_{f\in T_B\mathcal{F}}(\mu_{(X,Y)_{1:n}} - \lambda A_{(X,Y)_{1:n}})g_{f,1:n} \geq t\right).$$

Assuming that $B = 1/4$, which gives that $|g_f^k(x)| \leq 1$ for all k and x, the arguments in [BG19, Section 3.2.] lead to the inequalities

$$\begin{aligned}
&\mathbb{P}\left(\sup_{f\in T_{1/4}\mathcal{F}}(\mu_{(X,Y)_J^*} - \lambda A_{(X,Y)_J^*})g_{f,J} \geq t\right) \\
&\leq 3G_0(c)\mathbb{E}\left[\mathcal{N}_1(G_1(c,\lambda)t_0(c,\lambda,|J|), T_{1/4}\mathcal{F}, X_{1:n})\right]\exp(-b(c,\lambda)|J|t) \\
&\leq 3G_0(c)\left(\frac{e}{G_1(c,\lambda)t_0(c,\lambda,|J|)}\log(\frac{3e}{2G_1(c,\lambda)t_0(c,\lambda,|J|)})\right)^{V_{\mathcal{F}}}\exp(-b(c,\lambda)|J|t) \\
&=: a_0(c,\lambda,|J|)\exp(-b(c,\lambda)|J|t)
\end{aligned} \tag{3.13}$$

for every $J \subset \{1,\ldots,n\}$ and every random element $(X,Y)_J^*$ of $(\mathbb{R}^d\times\mathbb{R})^J$ with independent entries and the same marginals as $(X,Y)_J$, with G_0, G_1, and b given by

$$G_0(c) := 2(c+1)(2c+3), \qquad G_1(c,\lambda) := \frac{1}{8}\frac{1}{\lambda(c-1)+1}(1-\frac{1}{c}),$$

$$b(c,\lambda) := \frac{1}{2}\frac{1}{(\frac{1}{3}(1-\frac{1}{c})+(2\lambda-1)\frac{\lambda}{\lambda-1})^2}(1-\frac{1}{c})^3\frac{\lambda}{\lambda-1},$$

and provided that

$$t \geq t_0(c,\lambda,|J|) := \frac{-(\lambda-1)+\sqrt{(\lambda-1)^2+c(c+1)\lambda^2/|J|}}{2}.$$

Therefore we have, for every $J \subset \{1,\ldots,n\}$ and every $(X,Y)^*_J$ as indicated, the estimate

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}(\mu_{(X,Y)^*_J}-\lambda A_{(X,Y)^*_J})g_{f,J} \geq t\right) \leq \mathbf{1}_{\{t<t_0(|J|,c,\lambda)\}} + a_0(c,\lambda,|J|)\exp(-b(c,\lambda)|J|t)\mathbf{1}_{\{t_0(|J|,c,\lambda)\leq t\}}.$$

This gives rise, via Proposition 2.14 and elementary estimates, to the inequality

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}(\mu_{(X,Y)_{1:n}}-\lambda A_{(X,Y)_{1:n}})g_{f,1:n} \geq t\right) \leq (n\beta_{(X,Y)_{1:\infty}}(m)+L_{c,\lambda}(n,m,t))\mathbf{1}_{\{t\leq(1+\lambda)\}}, \tag{3.14}$$

where

$$L_{c,\lambda}(n,m,t) := \mathbf{1}_{\{t<t_0(\lfloor\frac{n}{m}\rfloor,c,\lambda)\}} + m\,a_0(c,\lambda,\left\lfloor\frac{n}{m}\right\rfloor+1)\exp(-b(c,\lambda)\left\lfloor\frac{n}{m}\right\rfloor t)))\mathbf{1}_{\{t_0(\lfloor\frac{n}{m}\rfloor,c,\lambda)\leq t.\}}$$

The desired estimate for the case $B = 1/4$ follows from (3.12) and integration with respect to t (and Lebesgue measure) of the right–hand side of (3.14), with the integral of $L_{c,\lambda}(n,m,\cdot)$ estimated as in the arguments following [BG19, Equation (3.13)]. The estimate for general $B > 0$ follows by an homogenization argument (see the homogenization argument after [BG19, Equation (3.19)]). □

3.4.3 Weak rates for $\beta-$mixing schemes

It is worth discussing what the right–hand side of (3.9) says about weak consistency, and to introduce some cases and consequences of special importance which fall under this discussion. The setting is again that in Section 3.4.1.

As a first consequence, we point out the following result:

Proposition 3.13 (Weak rate for uniformly bounded schemes)**.** *Assume that $\mathcal{F}$ is a pointwise measurable family with associated $VC-$dimension $V_{\mathcal{F}} < \infty$ and with*

$$\sup_{f,k}\{||f(X_k)||_{\mathbb{P},\infty}, ||Y_k||_{\mathbb{P},\infty}\} \leq B$$

for some $B \in (0,\infty)$, then for any sequence $(m_n)_n$ of natural numbers and any bounded positive sequence $(\delta_n)_n$,

$$\mathbb{E}\left[\mu_{X_{1:n}}|\widehat{\Phi}_n - \Phi_{1:n}|^2 - \inf_{f\in\mathcal{F}} \mu_{X_{1:n}}|f-\Phi_{1:n}|^2\right]$$
$$= O\left(\frac{\log n}{\delta_n n} m_n + n\beta_{(X,Y)_{1:\infty}}(m_n) + \delta_n\left(n\beta_{(X,Y)_{1:\infty}}(m_n) + \inf_{f\in\mathcal{F}} \mu_{X_{1:n}}|f-\Phi_{1:n}|^2\right)\right). \tag{3.15}$$

Proof. This is an immediate cosequence of (3.9) and (3.11), by choosing $B_n = B, V_{\mathcal{F}_n} = V_{\mathcal{F}}$, $\lambda_n = 1+\delta_n$ and (say) $c_n = 2$. □

Remark 3.14 (Some consequences of (3.15))**.** It follows in particular that, under the hypotheses of Proposition 3.13, the left-hand side of (3.15) converges to zero if there exists a sequence $(m_n)_n$ such that

$$m_n \frac{\log n}{n} + n\beta_{(X,Y)_{1:\infty}}(m_n) \to_n 0 \tag{3.16}$$

(take $\delta_n := (m_n \log n/n)^{1/2}$). Notice also that the rate at the right–hand side of (3.15) admits convenient interpretations in interesting cases: if for instance $(X,Y)_{1:\infty}$ is $m-$dependent (see item 3. in Section 2.6) and conditionally stationary in the sense that for some $\Phi : \mathbb{R}^d \to \mathbb{R}$, $\Phi(X_k) = \mathbb{E}[Y_k|X_k]$, $\mathbb{P}-$a.s., and if $\Phi \in \mathcal{F}$ (unbiased case), (3.15) gives the rate of convergence $O(\log n/n)$ to zero for the expected squared error of the least–squares loss $\mathbb{E}\left[\int_{\mathbb{R}^d} |\widehat{\Phi}_n(x) - \Phi(x)|^2 \,\mathrm{d}x\right]$ (take $m_n = m+1$ and $\delta_n = 1$ in (3.15)).

In any case (3.16) requires that

$$\beta_{(X,Y)_{1:\infty}}(m_n) = o(n^{-1}), \tag{3.17}$$

for some sequence $(m_n)_n$ satisfying $m_n = o(n/\log n)$. It is necessary for (3.16) that

$$\lim_n n\,\beta_{(X,Y)_{1:\infty}}(n) = 0 \tag{3.18}$$

because $(\beta_{(X,Y)_{1:\infty}}(n))_n$ is decreasing (see Remark 2.6), and in particular that

$$\lim_n \beta_{(X,Y)_{1:\infty}}(n) = 0. \tag{3.19}$$

Remark 3.15 (The "$\beta-$mixing" assumption)**.** Notice that, in general, (3.19) is less restrictive[19] than *the $\beta-$mixing assumption* on $(X,Y)_{1:\infty}$, which amounts to the hypothesis

$$\lim_m \sup_k \beta(\sigma((X,Y)_{1:k}), \sigma((X,Y)_{k+m:\infty})) = 0. \tag{3.20}$$

However, as we pointed out after Definition 2.8, (3.19) is exactly the beta-mixing assumption (3.20) when $(X,Y)_{1:\infty}$ is a Markov process.

For the rates in Definitions 2.7 and 2.8, we deduce the following versions of Theorem 3.10:

Proposition 3.16 (Weak rate of convergence for subexponentially $\beta-$mixing samples)**.** *There exists a universal constant C with the following property: if $(X,Y)_{1:\infty}$ is subexponentially β–mixing (Definition 2.7) with parameters (a,b,γ), and if for some $B \in (0,\infty)$ and all $k \in \{1,\dots,n\}$, $||Y_k||_{\mathbb{P},\infty} \leq B$, then for any*

$$1 < \lambda \leq \frac{3+\sqrt{1+8\sqrt{71}}}{4},$$

the statistical error in (3.9) is bounded by

$$\frac{C}{n}\left(\frac{B^2 V_{\mathcal{F}}}{(\lambda-1)}(1+\log n)+a\right)\left(\frac{2\log n}{b}\right)^{1/\gamma}$$

provided that

$$1 \leq \left(\frac{2\log n}{b}\right)^{1/\gamma} \leq \frac{n}{2}. \tag{3.21}$$

Proof. For any positive real number α and any $x \geq 2$, the inequality

$$\alpha x + an\exp(-b\lfloor x \rfloor^{\gamma}) \leq \alpha x + an\exp(-\frac{b}{2^{\gamma}}x^{\gamma})$$

holds. It follows that, if $C_{c',\lambda'}$ is the constant from (3.10) corresponding to $(c',\lambda') = (\sqrt{71}, 3+\sqrt{1+8\sqrt{71}}/4)$ and

$$\alpha_{\lambda,n} := 2C_{c',\lambda'}\frac{B^2 V_{\mathcal{F}}(\log\sqrt{71}+\log n)}{(\lambda-1)n}$$

[19] See the footnote on the definition of $\beta_{Z.}$ in page 14.

then, under the subexponentially mixing hypothesis (2.13), the statistical error in (3.9) is bounded by

$$\min_{x\in[2,n]}\{\alpha_{\lambda,n}x + an\exp(-\frac{b}{2^\gamma}x^\gamma)\}$$

Taking $x := 2^{1+1/\gamma}\left(\log n/b\right)^{1/\gamma}$, which lies in $[2,n]$ in virtue of (3.21), we get the bound

$$2^{1+1/\gamma}\left(\frac{\log n}{b}\right)^{1/\gamma}\alpha_{\lambda,n} + \frac{a}{n}.$$

for the statistical error in (3.9). The result follows from an easy estimation on this bound. □

A similar (and easier) argument, taking this time $x := \left\lceil n^{2/\gamma+1}\right\rceil$ and estimating via (2.14), gives the corresponding weak rate for subpolinomially $\beta-$mixing samples:

Proposition 3.17 (Weak rate of convergence for subpolinomially $\beta-$mixing samples)**.** *There exists a universal constant C with the following property: if $(X,Y)_{1:\infty}$ is subpolinomially β–mixing (Definition 2.8) with parameters (a,γ), and if for some $B\in(0,\infty)$ and all $k\in\{1,\dots,n\}$, $||Y_k||_{\mathbb{P},\infty}\leq B$, then for any*

$$1<\lambda\leq\frac{3+\sqrt{1+8\sqrt{71}}}{4},$$

the statistical error in (3.9) is bounded by

$$\frac{C}{n^{(\gamma-1)/(\gamma+1)}}\left(\frac{B^2V_{\mathcal{F}}}{(\lambda-1)}(1+\log n)+a\right).$$

Notice that, in Corollaries 3.16 and 3.17, we recover the rates of the independent case by letting $\gamma\to\infty$.

References

[Ada08] R. Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electronic Journal of Probability*, 13:1000–1034, 2008.

[Ber79] H.C.P Berbee. *Random walks with stationary increments and renewal theory.* Number 112 in Math. Centre Tracts. Matematisch Centrum, Amsterdam, 1979.

[BG19] D. Barrera and E. Gobet. Quantitative bounds for concentration-of-measure inequalities and empirical regression: the independent case. *Journal of Complexity*, 52:45–81, 2019.

[Bra05] R. Bradley. Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2:107–144, 2005.

[DDL+07] J. Dedecker, P. Doukhan, G. Lang, J.R. León, S. Louhichi, and C. Prieur. *Weak Dependence: with Examples and Applications*. Lecture Notes in Statistics. Springer, 2007.

[DFG09] R Douc, G Fort, and A Guillin. Subgeometric rates of convergence of f-ergodic strong Markov processes. *Stochastic Processes and their Applications*, 119(3):897–923, March 2009.

[DFMS04] R. Douc, G. Fort, E. Moulines, and P. Soulier. Practical drift conditions for subgeometric rates of convergence. *The Annals of Applied Probability*, 14(3):1353–1377, 2004.

[DG15] J. Dedecker and S. Gouëzel. Subgaussian concentration inequalities for geometrically ergodic Markov chains. *Electronic Communications in Probability*, 20, 2015.

[DMPS18] Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. *Markov Chains*. Springer Series in Operations Research and Financial Engineering. Springer, 2018.

[DMPS19] R. Douc, E. Moulines, P. Priouret, and P. Soulier. *Markov Chains*. Springer, to appear, 2019.

[DN93] A Dembo and A.B Nobel. A note on uniform laws of averages for dependent processes. *Statistics and Probability letters*, 17(3):169–172, 1993.

[Dou12] P. Doukhan. *Mixing: properties and examples*. Springer Science & Business Media, 2012.

[FGM17] G. Fort, E. Gobet, and E. Moulines. MCMC design-based non-parametric regression for rare-event. Application to nested risk computations. *Monte Carlo Methods and Applications*, 23(1):21–42, 2017.

[FM03a] G. Fort and E. Moulines. Convergence of the Monte Carlo expectation maximization for curved exponential families. *Annals of Statistics*, pages 1220–1259, 2003.

[FM03b] G. Fort and E. Moulines. Polynomial ergodicity of Markov transition kernels. *Stochastic Processes and their Applications*, 103(1):57–99, 2003.

[GKKW02] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics, 2002.

[JR02] S Jarner and G Roberts. Polynomial convergence rates of Markov chains. *The Annals of Applied Probability*, 12(1):224–247, 2002.

[KM17] V. Kuznetzov and M. Mohri. Generalization bounds for non-stationary mixing processes. *Machine Learning*, 106(1):93–117, January 2017.

[LT13] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[MPR09] F. Merlevède, M. Peligrad, and E. Rio. Bernstein inequality and moderate deviations under strong mixing conditions. In *High dimensional probability V: the Luminy volume*, pages 273–292. Institute of Mathematical Statistics, 2009.

[MR08] M Mohri and A Rostamizadeh. Rademacher complexity bounds for non i.i.d. processes. *Advances in neural information processing systems (NIPS)*, 21, 2008.

[MT09] S. Meyn and R.L. Tweedie. *Markov chains and stochastic stability*. Cambridge University Press, Cambridge, second edition, 2009.

[Nev75] J. Neveu. *Discrete parameter martingales*. North-Holland Publishing Company, Amsterdam-Oxford. American Elsevier Publishing Company, INC, New York, 1975.

[Pol90] D. Pollard. *Empirical Processes: Theory and Applications*. Institute of Mathematical Statistics and American Statistical Association, 1990.

[RM10] Q. Ren and M. Mojirsheibani. A note on nonparametric regression with β-mixing sequences. *Communications in Statistics-Theory and Methods*, 39(12):2280–2287, 2010.

[RST15] A Rakhlin, K Sridharan, and A Tewari. Sequential complexities and uniform martingale laws of large sequential complexities and uniform martingale laws of large numbers. *Probability theory and related fields*, 161(1–2):111–153, 2015.

[Sau72] N Sauer. On the density of families of sets. *Journal of combinatorial theory*, 13(1):145–147, July 1972.

[She72] S Shelah. A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific journal of mathematics*, 41(1):247–261, 1972.

[Son92] E. Sontag. Feedforward nets for interpolation and classification. *Journal of computer and system sciences*, 45:20–48, 1992.

[TT96] P Tuominen and R Tweedie. Subgeometric rates of convergence of f-ergodic Markov chains. *Adv. Appl. Prob.*, 26:775–798, 1996.

[VC71] V.N. Vapnik and Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, XVI(2):264–280, 1971.

[Vie97] G. Viennet. Inequalities for absolutely regular sequences: application to density estimation. *Probability theory and related fields*, 107(4):467–492, 1997.

[vW96] A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer Series in Statistics. New York, NY: Springer, 1996.

[Yu94] B Yu. Rates of convergence for empirical processes of stationary mixing sequences. *Annals of Probability*, 22(1):94–116, 1994.

[Yuk86] J.E. Yukich. Rates of convergence for classes of functions: the non i.i.d. case. *Journal of Multivariate Analysis*, 20(2):175–189, 1986.