Quantitative bounds for concentration-of-measure inequalities and empirical regression: the independent case

David Barrera, Emmanuel Gobet

Quantitative bounds for concentration-of-measure inequalities and empirical regression: the independent case*

David Barrera [†‡] Emmanuel Gobet [§]

Abstract

This paper is devoted to the study of the deviation of the (random) average L^2-error associated to the least–squares regressor over a family of functions $\mathcal{F}_n$ (with controlled complexity) obtained from n independent, but not necessarily identically distributed, samples of explanatory and response variables, from the minimal (deterministic) average L^2-error associated to this family of functions, and to some of the corresponding consequences for the problem of consistency.

In the i.i.d. case, this specializes as classical questions on least–squares regression problems, but in more general cases, this setting permits a precise investigation in the direction of the study of nonasymptotic errors for least–squares regression schemes in nonstationary settings, which we motivate providing background and examples.

More precisely, we prove first two nonasymptotic deviation inequalities that generalize and refine corresponding known results in the i.i.d. case. We then explore some consequences for nonasymptotic bounds of the error both in the weak and the strong senses. Finally, we exploit these estimates to shed new light into questions of consistency for least–squares regression schemes in the distribution–free, nonparametric setting.

As an application to the classical theory, we provide in particular a result that generalizes the link between the problem of consistency and the Glivenko–Cantelli property, which applied to regression in the i.i.d. setting over non–decreasing families $(\mathcal{F}_n)_n$ of functions permits to create a scheme which is strongly consistent in L^2 under the sole (necessary) assumption of the existence of functions in $\cup_n \mathcal{F}_n$ which are arbitrarily close in L^2 to the corresponding regressor.

KEYWORDS: empirical processes, concentration inequalities, empirical regression, distribution-free estimates, uniform deviation probability, consistency

MSC2010: 60F10, 60F15, 62GXX, 65C60, 68Q32

Contents

*The authors thank Gersende Fort for her useful feedbacks on this work.

†Email: jdbarrer@gmail.com, david.barrera@polytechnique.edu. CMAP, Ecole Polytechnique, Université Paris-Saclay, Route de Saclay, 91128 Palaiseau cedex, France. Supported by Chaire Marchés en Mutation, Fédération Française de Banque, and Institut Louis Bachelier.

‡Corresponding author

§Email: emmanuel.gobet@polytechnique.edu. CMAP, Ecole Polytechnique, Université Paris-Saclay, Route de Saclay, 91128 Palaiseau cedex, France. This research is part of the Chair *Financial Risks* of the *Risk Foundation*, the *Finance for Energy Market Research Centre* (FiME lab, from *Institut Europlace de Finance*), the Chair *Stress Test, RISK Management and Financial Steering* of the *Foundation Ecole polytechnique*.

1 Introduction

In this section, we introduce in a general manner the problem of this paper. We also present a short summary of the background literature in these topics, and we introduce the notation to be used along the discussions that follow.

1.1 Description of the problem

We depart from a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ supporting a sequence of random variables $(X_k, Y_k)_{k\in\mathbb{N}}$ taking values in $\mathbb{R}^d \times \mathbb{R}$ (and all the other random variables that will appear in our proofs). We assume that each Y_k is square-integrable and we denote by $\Phi_k : \mathbb{R}^d \mapsto \mathbb{R}$ a regression function of Y_k given X_k, so that Φ_k is a (measurable) function square integrable with respect to the law of X_k which satisfies

$$\Phi_k(X_k) = \mathbb{E}\left[Y_k \mid X_k\right] \tag{1.1}$$

$\mathbb{P}-$a.s.

Given an approximation space $\mathcal{F}$ of measurable functions from $\mathbb{R}^d$ to $\mathbb{R}$, for any $n \geq 1$, we define the empirical regression function $\widehat{\Phi}_n$ by

$$\widehat{\Phi}_n \in \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{k=1}^{n} |f(X_k) - Y_k|^2, \tag{1.2}$$

which stands in a way as a mean approximation of the regression functions Φ_k for $k = 1, \ldots, n$ ($\widehat{\Phi}_n$ might not be unique). Indeed, if we consider (1.2) with an expectation, by setting

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

then

$$\Phi_n^\star \in \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}|f(X_k) - \Phi_k(X_k)|^2,$$

which shows that $\widehat{\Phi}_n$ will presumably approximate simultaneously all individual regression functions Φ_k, $k = 1, \dots, n$.

While usually, in the literature, only the i.i.d. case for $(X_k, Y_k)_{k\in\mathbb{N}}$ is investigated (so that all Φ_k's are the same), here we focus on situations where the distributions may change with k, see Section 1.2 for a relevant application. Our aim is to investigate the statistical fluctuations arising between steps (1.2) and (1.3), this is tightly related to concentration-of-measure estimates uniformly on a class of functions.

This study about statistical fluctuations takes the form of how close are $\Phi_n^\star$ and $\widehat{\Phi}_n$ in various norms: for instance,

- Controlling the probability of large quadratic error $\mathbb{P}\left(\frac{1}{n}\sum_{k=1}^n \int_{\mathbb{R}^d} |\widehat{\Phi}_n(x) - \Phi_k(x)|^2 \mathbb{P}_{X_k}(\mathrm{d}x) > \varepsilon\right)$ where $\mathbb{P}_{X_k}(\mathrm{d}x)$ denotes the probability measure associated to X_k;
- Controlling the mean integrated squared error: $\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^n \int_{\mathbb{R}^d} |\widehat{\Phi}_n(x) - \Phi_k(x)|^2 \mathbb{P}_{X_k}(\mathrm{d}x)\right]$.

1.2 An example motivating the non i.i.d. case

In the case of i.i.d. data $(X_k, Y_k)_{k\in\mathbb{N}}$, this is the usual non-parametric regression problem with independent and identically distributed sample points (see [GKKW02, HTF09]). In this work, our concerns are related to the case where the i.i.d. assumption does not necessarily hold. Two applications serve to motivate this investigation:

(i) The first one is that of independent Monte Carlo schemes where (X_k, Y_k) is an approximation of some (X, Y) and a convergence at the limit holds as $k \to +\infty$; in that case we still have the independence but the sequence is not stationary anymore.

(ii) The second example is when the sequence describes a Markovian evolution, see the applications for nested risks computations as designed in [FGM17]. This part is developed in another work in progress, which builds over the results presented here.

Let us present an example about (i) of independent but non identically distributed samples: consider that a stochastic system is modelled by a Stochastic Differential Equation (SDE for short)

$$\mathrm{d}Z_t = b(t, Z_t)\mathrm{d}t + \sigma(t, Z_t)\mathrm{d}W_t, \quad Z_0 = z_0 \text{ deterministic}$$

with some drift and diffusion coefficients that are smooth enough as to ensure existence and uniqueness of the solution (typically Lipschitz continuous). For applications of SDE modelling, see [Oks00, KP10] for instance. Say that, in order to analyse the system, our aim is to compute:

$$\mathbb{E}\left[\varphi_1(Z_1)\right], \quad \mathbb{E}\left[\varphi_2(Z_2)\right], \quad \mathbb{E}\left[\varphi(Z_2) \mid Z_1\right],$$

for some functions $\varphi_1, \varphi_2, \varphi$ depending on the context of the application . A standard numerical method to compute the first two expectations is to use a Monte Carlo method which is made of three steps:

1. Discretize the time interval with a time-step $\Delta = 1/K$ (with $K \in \mathbb{N}$ fixed) and define the Euler scheme Z^Δ by (see [KP10])

$$Z_0^\Delta = z_0, \qquad Z_{(i+1)\Delta}^\Delta = Z_{i\Delta}^\Delta + b(i\Delta, Z_{i\Delta}^\Delta)\Delta + \sigma(i\Delta, Z_{i\Delta}^\Delta)(W_{(i+1)\Delta} - W_{i\Delta}) \text{ for } i \geq 0;$$

2. sample many independent copies of Z^Δ, which we denote $(Z^{\Delta,m} : 1 \leq m \leq M)$;

3. as a result, use the approximations

$$\mathbb{E}\left[\varphi_1(Z_1)\right] \approx \frac{1}{M}\sum_{m=1}^{M} \varphi_1(Z_1^{\Delta,m}), \quad \mathbb{E}\left[\varphi_2(Z_2)\right] \approx \frac{1}{M}\sum_{m=1}^{M} \varphi_2(Z_2^{\Delta,m}). \tag{1.4}$$

By tuning optimally $\Delta \to 0$ and $M \to +\infty$, we can show that this method has an order of convergence equal to $1/3$ regarding the computational cost $\mathcal{C}$, i.e. the error tolerance decreases (under mild assumptions, see [DG95, Theorem 1, with $p = q = 1$] for details) as $\mathcal{C}^{1/3}$. Although simple to implement, this method is not optimal.

A more efficient method is to use Multi-Level Monte Carlo (MLMC for short) methods. This methodology was initiated by Heinrich [Hei01] and Giles [Gil08], and it has been strongly developed in the last decade, see [Gil15] for a review. In its native form of [Gil08], the MLMC estimator writes

$$\mathbb{E}\left[\varphi_1(Z_1)\right] \approx \frac{1}{M_0}\sum_{m=1}^{M_0} \varphi_1(Z_1^{\Delta_0,0,m}) + \sum_{l=1}^{L} \frac{1}{M_l}\sum_{m=1}^{M_l} \left(\varphi_1(Z_1^{\Delta_l,l,m}) - \varphi_1(Z_1^{\Delta_{l-1},l,m})\right), \tag{1.5}$$

and similarly for estimating $\mathbb{E}\left[\varphi_2(Z_2)\right]$. The simulations at a given level $l \in \{0, \dots, L\}$, i.e. $(Z^{\Delta_l,l,m} : 1 \leq m \leq M_l)$ are i.i.d. with same distribution as Z^{Δ_l}, and the simulations across levels $l = 0, \dots, L$ are independent.

The heuristics behind this method is to provide a rough approximation with the simulations from level $l = 0$ (by taking Δ_0 not small), and then to provide corrections with smaller and smaller variances (the difference $\varphi_1(Z_1^{\Delta_l,l,m}) - \varphi_1(Z_1^{\Delta_{l-1},l,m})$ reads like a control variate). Upon chosing appropriately the numerical parameters M_l and Δ_l to find the optimal trade-off between bias and variance, it can be shown that the estimator (1.5) achieves a order of convergence of $1/2$ (which is optimal when using a Monte-Carlo method) under mild assumptions, see [Gil15] for details.

Now it remains to evaluate the conditional expectation function $z \mapsto \mathbb{E}\left[\varphi(Z_2) \mid Z_1 = z\right]$.

- With the standard Monte-Carlo procedure (see (1.4)), we can perform a empirical regression using the i.i.d. samples $(Z_1^{\Delta,m}, Z_2^{\Delta,m} : 1 \leq m \leq M)$: this is the usual i.i.d. setting for regression.
- With the MLMC scheme (see (1.5)), in order to recycle samples and to avoid wasting simulations, we would like to use the non i.i.d samples $(Z_1^{\Delta_l,l,m}, Z_2^{\Delta_l,l,m} : 1 \leq m \leq M_l, 1 \leq l \leq L)$. The question that pops up is how does this impact the error estimates on the regression function.

The goal of this work is to develop some new tools, in particular able to quantify the regression error in a non i.i.d. setting, like in this MLMC context. Further applications will be developed in subsequent works.

1.3 Background results in the literature

The study of uniform large deviations for empirical means and their consequences in the i.i.d. case is arguably the central topic in the theory of the Empirical Process and therefore has a very rich history, an account of which can be found in [VW96, Notes to chapter 2] and the references therein.

The technical developments in this paper have their origin, in a considerable part, on the exposition about distribution–free rates of convergence for least–squares regression schemes with i.i.d. samplings presented in [GKKW02, Chapter 11], which uses also some results from the previous exposition on uniform law of large numbers in [GKKW02, Chapter 9]. The notes to these two chapters give a satisfactory account of references describing the evolution of some of the techniques at the heart of our proofs.

Of special mention among them is the idea of symmetrization, which dates at least back to 1971 with the seminal work [VC71] by Vapnik and Chervonenkis on the uniform convergence of relative frequencies of events to their probabilities, a work aimed to extend the classical Glivenko-Cantelli Theorem [Gli33, Can33] to more general sets of indicator functions. According to [VW96, p. 270], these ideas appeared already in 1968 in the work [Kah68] by Kahane. The ideas by Vapnik and Chervonenkis have grown up to constitute a statistical learning theory on its own, see [Vap00] for a systematic introduction.

Consistency is of course a major issue in the research on any method of statistical inference. See [GKKW02] for a collection of consistency results related to nonparametric least–squares regression estimates in the i.i.d. case. For general results in this direction under the additional assumption of independent errors, see [GW96]. For a recent study on convergence rates see [HW17] and the references therein.

The independent, not necessarily identically distributed (i.n.n.i.d for short) case, by contrast, has not been explored on its own with nearly such intensity, presumably due, at the level of applications, to the overwhelming importance that the i.i.d. case has for many of the practical problems that have challenged the statisticians and, at the theoretical level, to the ambiguous meaning of the empirical mean as an approximation of an integral in the case in which the sampling sequence is nonstationary.

Now, the seminal inequalities of Bernstein [Ber24] and Hoeffding [Hoe63], which are fundamental in what follows and appear in many results on the Empirical Process, belong to the i.n.n.i.d. case. Note also that all the large deviations and ergodic theoretic literature for nonstationary sequences intersects the i.n.n.i.d. case (for instance: every independent sequence is also a Markov chain, for which the "i.d." hypothesis is equivalent to homogeneity). We stress out also the fact that, thanks to the existence of "coupling" techniques (see [PG99] and [DDL+07] for expositions on these topics), the study of problems under the i.n.n.i.d. hypothesis is of importance at the theoretical level beyond its own boundaries. This later idea will be exploited in a continuation of this paper.

The i.n.n.i.d. case has nonetheless been considered in diverse scenarios in the theory of large deviations and of the Empirical Process. [Ben62] is aimed to improve the bounds in Bernstein's inequality. [Wel81] develops a Glivenko-Cantelli theorem under the Prohorov and dual-bounded-Lipschitz metrics for the i.n.n.i.d. case under the hypothesis of tightness for the average measures. [Zui78] extends and refines some deviation inequalities between the empirical and the "true" distribution functions in the i.i.d. case to the corresponding inequalities in the i.n.n.i.d. case with respect to the average of the true distribution functions. Notice that all these results deal with concentration of measure inequalities either for finite sets of functions or indicator functions parametrized by the real line, whereas our results hold in a more general setting.

In the asymptotic setting (our deviation results are nonasymptotic), we mention [Pol90, Theorem 10.6] [1], which is a functional central limit theorem for empirical processes in the i.n.n.i.d. context, a development continued by [Kos03] in the direction of multiplier central limit theorems. A motivation by example of the relevance of this case is discussed also in [Kos03].

As an instance of this topic in the realm of applied learning, see [MM12], where generalization bounds in terms of the Rademacher complexity are derived for i.n.n.i.d. samplings.

We observe finally that in the learning and forecasting applications, nonstationarity (specially under Markovian assumptions) is already an important topic of research and, as already pointed out, the results developed in the i.n.n.i.d. case can be extended to more general nonstationary settings via coupling techniques. We do not give an account here on the developments of these topics for dependent, nonstationary designs because they escape the specific purpose of the present paper. For an exposition in this direction, see [KM17] and the references therein.

Our contribution. The results presented in this paper are of two kinds: *uniform concentration inequalities for the empirical process* and *consistency theorems for least squares regression estimates*, both within the context of VC-subgraph families of functions[2] in the i.n.n.i.d. case.

Our results, while valid for the case of i.n.n.i.d. sampling, improve also the current state of the art (as far as the authors know) in the classical i.i.d. context. Summarized, our contributions follow the following three directions:

1. *Extension of VC-methods to more general independent cases.* As the reader shall see, the proofs of Theorems 2.2 and 2.3 proceed by generalizing the VC methods available for the i.i.d. case. The results obtained open the way to new applications and, as a continuation of this paper will show, have remarkable consequences for the theory in (dependent) cases for which the associated samplings are nonstationary. The results obtained also permit to recover their i.i.d. counterparties without any loss.

2. *Refined estimates.* Indeed, our conclusions refine dramatically the corresponding i.i.d. results which inspired them. This is due to the fact that we are able to obtain "parametrized" deviation inequalities (see the estimates (2.2), (2.3) and (2.6)) in a way that allows for a better optimization of constants and which, even more, permits to study their consequences using diagonal arguments, thus opening the exploration of a new realm of convergence results. These results permit also a refined investigation of convergence rates and of nonasymptotic estimates. See for instance Theorems 3.1 and 3.10.

3. *Novel consistency results.* As an instance of the gains obtained in asymptotics, we provide a series of weak and strong consistency results for least–squares regression schemes, see Section 3.3. In particular, we provide a least–squares regression scheme for the i.i.d. case (Theorem 3.14) which is strongly consistent under hypotheses that generalize several schemes previously studied in a case–by–case basis and which does not require the Glivenko-Cantelli property as an assumption.

An important message of our results is that, by considering average measures, the nonasymptotic deviation bounds obtained in the i.i.d. case via symmetrization methods can be extended to the i.n.n.i.d.

[1] Indeed, this book works under a particular presentation of the i.n.n.i.d. assumption.
[2] But see Remarks 3.3, 3.5 and 3.19 below.

case without additional assumptions. This idea, while not new (see for instance the results in [Wel81] and [Zui78], and of course [Pol90]), seems to have not been elaborated in the context of regression methods, and we hope that it will serve as an inspiring point for the extension of many of the results within the learning theory for i.i.d. samples to nonstationary settings. Besides of the interest that this idea might have of its own, it will be evident in a continuation of this paper that it helps to fill some of the gaps in learning theory for methods that rely on nonstationary (and possibly dependent) samplings (see [FGM17] for an application in which the design points $(X_k)_{k\in\mathbb{N}}$ come from MCMC methods).

1.4 Notation and organisation of the paper

Besides the notation already introduced in Section 1.1, we will use the following conventions

i) We frequently use the notation $w_{1:n}$ for a (finite) sequence $w_1, \ldots, w_n$.

ii) If W is a random variable defined on $(\Omega, \mathcal{A}, \mathbb{P})$ and $B \subset \mathbb{R}$ is a Borel set, $\{W \in B\}$ denotes the $\mathcal{A}-$set given by $\{W \in B\} := \{\omega \in \Omega : W(\omega) \in B\}$.

iii) For a sample $D_n = ((X_k, Y_k))_{k=1}^n$ of the random variables in Section 1.1, denote by $\mathbb{E}_n$ and $\widetilde{\mathbb{E}}_n$ the conditional expectations with respect to $X_{1:n}$ (the "explanatory sample") and D_n (the "full experiment")

$$\mathbb{E}_n W := \mathbb{E}\left[W \mid X_1, \ldots X_n\right], \qquad \widetilde{\mathbb{E}}_n W := \mathbb{E}\left[W \mid (X_1, Y_1), \ldots, (X_n, Y_n)\right], \tag{1.6}$$

for every $\mathbb{P}-$integrable W. Conditional probabilities $\mathbb{P}_n$ and $\widetilde{\mathbb{P}}_n$ are defined similarly (use indicator functions in (1.6)).

iv) $\mathcal{F}$ is a class of measurable functions from $\mathbb{R}^d$ to $\mathbb{R}$, such that for any $f \in \mathcal{F}$, $f(X_k)$ is in $L^2_{\mathbb{P}}$, for any k. We will use the VC-dimension associated to $\mathcal{F}$, $V_{\mathcal{F}}$, defined as the Vapnik-Chervonenkis dimension of the family of sets

$$\{\{(x, y) \in \mathbb{R}^d \times \mathbb{R} : y \leq f(x)\} : f \in \mathcal{F}\}$$

as a measure of the complexity of $\mathcal{F}$, see [GKKW02, Definition 9.6]. In simple cases, $\mathcal{F}$ has the structure of a finite-dimensional vector space over $\mathbb{R}$ with finite dimension, and then the bound

$$V_{\mathcal{F}} \leq \dim(\mathcal{F}) + 1$$

holds (see [GKKW02, p.152]).

v) We will use boldface characters to denote sequences of functions with the same domain and co–domain. Typically $\mathbf{g} = (g_k)_k$ where each $g_k : \mathbb{R}^d \to \mathbb{R}$ is a Borel measurable function.

vi) L_1-*Covering Numbers.* Given a sample $z_{1:n}$, and using the semi-norm

$$|\mathbf{g}_1 - \mathbf{g}_2|_{z_{1:n},1} := \frac{1}{n}\sum_{k=1}^{n} |g_{1,k}(z_k) - g_{2,k}(z_k)|$$

(where $\mathbf{g}_j := (g_{j,k})_k$, $j = 1, 2$), consider a δ-covering $\mathbf{g}^1, \ldots, \mathbf{g}^M$ of the family of sequences of functions $\mathcal{G} = \{\mathbf{g} = (g_k)_k\}$: for any $\mathbf{g} \in \mathcal{G}$, there is a $\mathbf{g}^i$ such that $|\mathbf{g} - \mathbf{g}^i|_{z_{1:n},1} \leq \delta$. The minimal of such M's is the L_1-δ-covering number of $\mathcal{G}$ at $z_{1:n}$, and is denoted $\mathcal{N}_1(\delta, \mathcal{G}, z_{1:n})$. We will refer to the (eventually random) semi-norm $|\cdot|_{z_{1:n},1}$ as the *empirical* L_1 *norm at* $z_{1:n}$.

vii) We write $T_B : \mathbb{R} \mapsto \mathbb{R}$ for the soft truncation operator at the level $B \geq 0$:

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

viii) To simplify notation for averages and expectations, we introduce the operators A_n, $\widetilde{A}_n$, μ_n, $\widetilde{\mu}_n$ averaging sequences of functions over the samples (the empirical measure) or over the exact distributions:

$$A_n \mathbf{f} = \frac{1}{n} \sum_{k=1}^{n} f_k(X_k), \qquad \mu_n \mathbf{f} = \frac{1}{n} \sum_{k=1}^{n} \int_{\mathbb{R}^d} f_k(x) \mathbb{P}_{X_k}(\mathrm{d}x), \tag{1.7}$$

$$\widetilde{A}_n \mathbf{f} = \frac{1}{n} \sum_{k=1}^{n} f_k(X_k, Y_k), \qquad \widetilde{\mu_n} \mathbf{f} = \frac{1}{n} \sum_{k=1}^{n} \int_{\mathbb{R}^d \times \mathbb{R}} f_k(x, y) \mathbb{P}_{X_k, Y_k}(\mathrm{d}x, \mathrm{d}y), \tag{1.8}$$

where $\mathbf{f} = (f_k)_{k \in 1, \ldots, n}$ is a sequence of (possibly random, depending on data) functions from $\mathbb{R}^d \to \mathbb{R}$ (or $\mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$). Here $\mathbb{P}_{X_k}(\mathrm{d}x)$ and $\mathbb{P}_{X_k, Y_k}(\mathrm{d}x, \mathrm{d}y)$ are the laws of X_k and (X_k, Y_k) respectively.

Organisation of the paper. First, in Section 2, we derive general concentration-of-measure inequalities for independent but non i.i.d. sequences. Then in Section 3, we apply them to regression estimates and establish asymptotic consistency results.

2 Uniform concentration-of-measure inequalities

This section is devoted to the statement and proofs of the two main deviation inequalities of this paper (Theorems 2.2 and 2.3). We start with a short motivation on the relationship between large deviations and expectations of a probability distribution, which is also an introduction to some of the applications to be developed in Section 3.

2.1 Nonasymptotic deviations (Theorems 2.2 and 2.3)

We will start by a tight deviation bound, uniformly on the class of functions, which is interesting on its own. This generalizes [GKKW02, Theorem 11.6], with tighter constants. See Sections 1.1 and 1.4 for notation.

Let us provide, as a preamble, some easy motivation: having better constants in exponential inequalities helps much to improve moment estimates. To illustrate this claim, start by considering the following lemma:

Lemma 2.1. *Let Z be an integrable real random variable such that for some constants $(a, b, t_0) \in (0, \infty) \times (0, \infty) \times [0, \infty)$ the estimate*

$$\mathbb{P}(Z \geq t) \leq a \exp(-bt)$$

holds for every $t \geq t_0$. Then

$$\mathbb{E}[Z] \leq (t_0 + \frac{a}{b} \exp(-bt_0)) \mathbf{1}_{[a < \exp(bt_0)]} + \frac{1}{b}(1 + \log a) \mathbf{1}_{[a \geq \exp(bt_0)]}. \tag{2.1}$$

Proof. For every $t \geq t_0$

$$\mathbb{E}[Z] \leq \mathbb{E}[Z\mathbf{1}_{Z\geq 0}] = \int_0^\infty \mathbb{P}\left(Z \geq t'\right) \mathrm{d}t' \leq t + a\int_t^\infty \exp(-bt')\,\mathrm{d}t' = t + \frac{a}{b}\exp(-bt).$$

From here (2.1) follows via an elementary minimization argument. □

Thus for any constants $a \geq 1$ and $b > 0$, we have the implication, available for any non-negative random variable Z,

$$\mathbb{P}(Z > \varepsilon) \leq a\exp(-bn\varepsilon), \forall \varepsilon > 0 \quad \Longrightarrow \quad \mathbb{E}[Z] \leq \frac{1}{bn}\big(1 + \log(a)\big).$$

Therefore, improving the constant a has only an impact as a logarithm, thus it is minor; on the other hand, deriving a larger b in the upper bound for $\mathbb{P}(Z > \varepsilon)$ significantly improves the final upper bound on $\mathbb{E}[Z]$. Our subsequent mathematical derivation is inspired by these kind of observations, in particular in order to get "the smaller factor b".

Theorem 2.2. *Let $B > 0$ and let $\mathcal{F}$ be a pointwise measurable*[3] *collection of sequences $\mathbf{f} = (f_k)_{k\in 1,\ldots,n}$ of Borel-measurable functions $\mathbb{R}^d \to [0, B]$. Then for every*

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

we have

$$\begin{aligned}
&\mathbb{P}\left(\exists \mathbf{f} \in \mathcal{F} : (1-\varepsilon)\frac{1}{n}\sum_{k=1}^n f_k(X_k) - (1+\varepsilon)\frac{1}{n}\sum_{k=1}^n \int_{\mathbb{R}^d} f_k(x)\mathbb{P}_{X_k}(\mathrm{d}x) > \alpha\varepsilon\right) \\
&\qquad \leq \frac{2\gamma}{\gamma-1}\mathbb{E}\left[\mathcal{N}_1(\frac{\alpha}{\gamma'}\frac{(c-1)}{(c-\varepsilon)}\frac{\varepsilon}{2}, \mathcal{F}, X_{1:n})\right]\exp\left(-\frac{2\varepsilon^2}{(2-(1+\frac{1}{c})\varepsilon)^2}(1-\frac{1}{c})^2(1-\frac{1}{\gamma'})\frac{\alpha}{B}n\right), \qquad (2.2)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}\left(\exists \mathbf{f} \in \mathcal{F} : (1-\varepsilon)\frac{1}{n}\sum_{k=1}^n \int_{\mathbb{R}^d} f_k(x)\mathbb{P}_{X_k}(\mathrm{d}x) - (1+\varepsilon)\frac{1}{n}\sum_{k=1}^n f_k(X_k) > \alpha\varepsilon\right) \\
&\qquad \leq \frac{2\gamma}{\gamma-1}\mathbb{E}\left[\mathcal{N}_1(\frac{\alpha}{\gamma'}\frac{(c-1)}{(c+c\varepsilon)}\frac{\varepsilon}{2}, \mathcal{F}, X_{1:n})\right] \times \exp\left(-\frac{2\varepsilon^2}{(2+(1+\frac{1}{c})\varepsilon)^2}(1-\frac{1}{c})^2(1-\frac{1}{\gamma'})\frac{\alpha}{B}n\right), \qquad (2.3)
\end{aligned}$$

provided that

$$n \geq \frac{B^2\gamma}{4\alpha(\alpha+B)}\left(\frac{c}{\varepsilon}\right)^2.$$

The previous controls can take other forms when combined with the regression problems. This is an extension to the non i.i.d. case of [GKKW02, Theorem 11.4], with refined estimates, especially regarding the factor in the exponential term which plays usually an important role in the subsequent estimates

[3] I.e., such that there exists a countable subfamily $(\mathbf{f}_j)_j$ of $\mathcal{F}$ ($\mathbf{f}_j = (f_{j,k})_k$) with the property that, for every $\mathbf{f} = (f_k)_k \in \mathcal{F}$, there is a sequence $(j_l)_l$ such that, for every $1 \leq k \leq n$ and every $x \in \mathbb{R}^d$,

$$\lim_l f_{j_l,k}(x) = f_k(x).$$

(see later the discussion in Remark 3.2). Besides we correct an error in their proof (at line 6 from the top of [GKKW02, p.213], one should read $\{g_f^2, f \in \mathcal{F}\}$ instead of $\{g_f, f \in \mathcal{F}\}$), which changes slightly the upper bound. In what follows, we use the character "y" to denote the projection $\mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ on the second coordinate

$$y(x_0, y_0) = y_0. \tag{2.4}$$

Theorem 2.3. *Assume that $\mathcal{F}$ is a pointwise measurable family of functions $\mathbb{R}^d \to \mathbb{R}$, and that for some $B > 0$*

$$|Y_k - f(X_k)| + |Y_k - \Phi_k(X_k)| \leq B, \tag{2.5}$$

$\mathbb{P}-a.s.$ for all $f \in \mathcal{F}$ and all $k \in \{1, \dots, n\}$. Then for every $(\alpha, \varepsilon) \in (0, \infty) \times (0, 1)$ and every $(\rho, \gamma, \gamma', c) \in (0, 1) \times (1, \infty) \times (1, \infty) \times (1, \infty)$ we have, for $u = \pm 1$

$$\begin{aligned}
&\mathbb{P}\Bigg[\exists f \in \mathcal{F} : u\Big(\frac{1}{n}\sum_{k=1}^n \int_{\mathbb{R}^d \times \mathbb{R}} \Big(|f(x) - y|^2 - |\Phi_k(x) - y|^2\Big)\mathbb{P}_{X_k, Y_k}(\mathrm{d}x, \mathrm{d}y) \\
&\qquad - \frac{1}{n}\sum_{k=1}^n \Big(|f(X_k) - Y_k|^2 - |\Phi_k(X_k) - Y_k|^2\Big)\Big) > \\
&\varepsilon\Big(\alpha + \frac{1}{n}\sum_{k=1}^n \int_{\mathbb{R}^d \times \mathbb{R}} \Big(|f(x) - y|^2 - |\Phi_k(x) - y|^2\Big)\mathbb{P}_{X_k, Y_k}(\mathrm{d}x, \mathrm{d}y)\Big)\Bigg] \\
&\leq \left(2\frac{\gamma}{\gamma - 1}\right)^2 \mathbb{E}\left[\mathcal{N}_1(\frac{1}{8B}\frac{(c-1)}{(c-\varepsilon)}\frac{1}{\gamma'}\alpha\varepsilon, \mathcal{F}, X_{1:n})\right] \\
&\quad \times \exp\left(-\frac{1}{2B^2}\frac{1}{(1 - \frac{1}{2}(1 + \frac{1}{c})\varepsilon)^2}(1 - \frac{1}{c})^2(1 - \frac{1}{\gamma'})\varepsilon^2\alpha n\right) \\
&+ 2\frac{\gamma}{\gamma - 1}\mathbb{E}\left[\mathcal{N}_1(\frac{1}{4B}\frac{1}{(1 - \frac{1}{\gamma'})\varepsilon(1 - \varepsilon) + (1 + \varepsilon)}(1 - \frac{1}{\gamma'})\alpha\varepsilon\rho, \mathcal{F}, X_{1:n})\right] \\
&\quad \times \exp\left(-\frac{1}{2B^2}\frac{(1 - \varepsilon)}{(\frac{1}{3}(1 - \frac{1}{\gamma'})\varepsilon(1 - \varepsilon) + (1 + \varepsilon))^2}(1 - \frac{1}{\gamma'})^2(1 - \rho)\varepsilon^2\alpha n\right)
\end{aligned} \tag{2.6}$$

provided that

$$n \geq \frac{\gamma}{4\alpha(B^2 + \alpha)}\frac{B^4}{\varepsilon^2}\max\{\gamma'^2, c^2\}. \tag{2.7}$$

Besides of allowing for better estimates of the (generic) constants known so far in the i.i.d. case, the generality of the parameters $(\rho, \gamma, \gamma', c)$ appearing in the above deviation inequalities is very helpful to develop new or improved results. For instance, to get the smallest upper bound in the generalization error of Theorem 3.1, we should take c large, which in view of (3.7) corresponds to $\gamma' \to +\infty$, $\gamma \to 1$, $\rho \to 0$ as $n \to +\infty$ in an appropriate manner (see Remark 3.2). Also, for proving the consistency results of Section 3.3, we will take $\lambda \to 1$, $\alpha \to +\infty$, $\varepsilon \to 0$, according to some joint relations. Actually, for many applications, $(\rho, \gamma, \gamma', c, \alpha, \varepsilon)$ have to be chosen dependently to take advantage of the working assumptions.

Remark 2.4 (About the condition (2.5))**.** It is likely that choosing independently the data (X_k, Y_k) and the function space $\mathcal{F}$ does not ensure that (2.5) holds, since this condition bridges both. However,

this condition is quite flexible and allows in some cases to adjust $\mathcal{F}$ to (X_k, Y_k). Let us exemplify this: assume that for some constant $B > 0$ we know a function $F : \mathbb{R}^d \to \mathbb{R}$ (a "trend" for the data (X_k, Y_k), *which might be unbounded*) such that the properties

$$|Y_k - F(X_k)| \leq \frac{B}{4}, \qquad \sup_{f \in \mathcal{F}} |f(X_k)| \leq \frac{B}{4}. \qquad \mathbb{P} - a.s.,$$

hold for all $k \in \mathbb{N}$. Then the new set of functions $\widetilde{\mathcal{F}} = \{\widetilde{f} = F + f : f \in \mathcal{F}\}$ satisfies (2.5): indeed the conditional Jensen's inequality gives that (for any $k \in \mathbb{N}$)

$$|\Phi_k(X_k) - F(X_k)| = |\mathbb{E}\left[Y_k - F(X_k) \mid X_k\right]| \leq \frac{B}{4}, \qquad \mathbb{P} - a.s.,$$

and therefore we have that for all $\widetilde{f} \in \widetilde{\mathcal{F}}$

$$|Y_k - \widetilde{f}(X_k)| + |Y_k - \Phi_k(X_k)| \leq 2|Y_k - F(X_k)| + |\widetilde{f}(X_k) - F(X_k)| + |\Phi_k(X_k) - F(X_k)| \leq B, \quad \mathbb{P} - a.s..$$

2.2 Proof of Theorem 2.2

2.2.1 Preliminaries

Our proof is inspired from [GKKW02, Chapters 9-10-11], but with significant differences firstly to account for the non-stationarity of the sequence and secondly to obtain more refined estimates. Additionally we establish bilateral deviations, as a difference with [GKKW02, Chapters 9-10-11].

1. A first key ingredient in the analysis is to switch from the *original sample* $(X_k, Y_k)_k$ to a *ghost sample* $(X'_k, Y'_k)_k$, which is an independent copy of the initial sample.

2. A second key ingredient is to use concentration-of-measure inequalities for a fixed value n of the sample size, on suitable functions g_f that allow to bound the variance by the expectation up to a constant. Namely, for a given $f : \mathbb{R}^d \to \mathbb{R}$, set

$$g_f^k(x, y) := |f(x) - y|^2 - |\Phi_k(x) - y|^2, \tag{2.8}$$

where Φ_k is given by (1.1), and we denote by $\mathbf{g}_f$ the sequence of functions

$$\mathbf{g}_f := (g_f^k)_k. \tag{2.9}$$

We extend notations (1.7) and (1.8) for averages over the ghost sample:

$$A'_n \mathbf{f} = \frac{1}{n} \sum_{k=1}^{n} f_k(X'_k), \qquad \widetilde{A}'_n \mathbf{f} = \frac{1}{n} \sum_{k=1}^{n} f_k(X'_k, Y'_k),$$

for sequences of real valued functions $\mathbf{f} = (f_k)_k$.

Operations between sequences. To carry out the proofs below, we will use the "component-wise" operations between sequences (of -possibly random- functions): given $\mathbf{f}_1 = (f_{1,k})_k$ and $\mathbf{f}_2 = (f_{2,k})_k$, $\mathbf{f}_1 + \mathbf{f}_2 := (f_{1,k} + f_{2,k})_k$ and $\mathbf{f}_1 \mathbf{f}_2 = (f_{1,k} f_{2,k})_k$. We also operate with sequences and constants via $a\mathbf{f} = a(f_k)_k := (af_k)_k$ and $a + \mathbf{f} := (a + f_k)_k$.

It is important to point out that we will usually deal in the sequel with coverings that either depend on families $\mathcal{F}$ given by constant sequences of functions (each $\mathbf{f} \in \mathcal{F}$ is of the form $\mathbf{f} = (f, \ldots, f)$) or on families $\mathcal{G}_{\mathcal{F}}$ of sequences of functions $\mathbf{g}_f$ as in (2.9). This will keep our approximating families away from overfitting problems.

2.2.2 Proof of Inequalities (2.2) and (2.3)

We open the proofs of (2.2) and (2.3), which are basically identical, with the following general lemma. The notation is that in Section 1.4.

Lemma 2.5. *For fixed $\eta, \varepsilon > 0$ and any Borel-measurable $\mathbf{f} = (f_k)_{k\in 1,\dots,n}$ with $0 \le f_k(x) \le B$ (for all k and x) we have, for $u = \pm 1$*

$$\mathbb{P}\left(u(A_n\mathbf{f} - \mu_n\mathbf{f}) > \varepsilon(\eta + \mu_n\mathbf{f})\right) \le \frac{B^2}{4n\varepsilon^2\eta(\eta + B)}.$$

Proof. Assume first that $B = 1$. Apply the Chebyshev inequality, together with independence between X_k and X_l for $k \neq l$: by setting $m_k := \int_{\mathbb{R}^d} f_k(x)\mathbb{P}_{X_k}(\mathrm{d}x)$ and $m = \frac{1}{n}\sum_{k=1}^n m_k = \mu_n\mathbf{f}$, this gives

$$\begin{aligned}\mathbb{P}\left[\frac{u(A_n\mathbf{f} - \mu_n\mathbf{f})}{\eta + \mu_n\mathbf{f}} \ge \varepsilon\right] &\le \sum_{k=1}^n \frac{\int_{\mathbb{R}^d} f_k^2(x)\mathbb{P}_{X_k}(\mathrm{d}x) - \left(\int_{\mathbb{R}^d} f_k(x)\mathbb{P}_{X_k}(\mathrm{d}x)\right)^2}{n^2\varepsilon^2(\eta + \mu_n\mathbf{f})^2} \\ &\le \frac{1}{n\varepsilon^2}\frac{\frac{1}{n}\sum_{k=1}^n m_k - \frac{1}{n}\sum_{k=1}^n m_k^2}{(\eta + \frac{1}{n}\sum_{k=1}^n m_k)^2}\end{aligned}$$

(use the Jensen inequality for the average in k)

$$\le \frac{1}{n\varepsilon^2}\frac{m - m^2}{(\eta + m)^2}. \tag{2.10}$$

The above upper bound is maximal at $m = \frac{\eta}{2\eta+1}$ and the maximum is $\frac{1}{4n\varepsilon^2\eta(\eta+1)}$.

This gives the upper bound in the case $B = 1$. The general case $B > 0$ follows from dividing all the quantities inside the respective probabilities by B and applying the case $B = 1$. □

Proof of Inequality (2.2)

We have to bound

$$\mathbb{P}\left(\bigcup_{\mathbf{f}\in\mathcal{F}}\left\{A_n\mathbf{f} - \mu_n\mathbf{f} > \varepsilon(\alpha + A_n\mathbf{f} + \mu_n\mathbf{f})\right\}\right).$$

To facilitate a first reading, we have split the proof in a series of partial results: Lemmas 2.6 to 2.9. By looking at these it is easy to see what is exactly the chain of inequalities leading to (2.2) for the case $B = 1$, each of whose links can be verified separately reading the arguments behind these partial results. We have integrated these arguments in a manner that allows an alternative, "linear" reading of the proof of (2.2) (with $B = 1$). Finally, the conclusion for general $B > 0$ is given as a last step via an easy homogenization argument.

Assume thus that $B = 1$ to begin with, and let $(\alpha, \varepsilon, c, \gamma, \gamma')$ be as in the statement of the theorem.
▷ *Symmetrization.* Observe that for $\mathbf{f} \in \mathcal{F}$ the set

$$\{A_n\mathbf{f} - \mu_n\mathbf{f} > \varepsilon(\alpha + A_n\mathbf{f} + \mu_n\mathbf{f})\} \cap \{A'_n\mathbf{f} - \mu_n\mathbf{f} \le \frac{\varepsilon}{c}(\alpha + A'_n\mathbf{f} + \mu_n\mathbf{f})\},$$

is included, by substracting the inequalities describing each of the sets in the intersection, in the set

$$\left\{A_n\mathbf{f} - A'_n\mathbf{f} > \varepsilon\left((1 - \frac{1}{c})\alpha + A_n\mathbf{f} - \frac{1}{c}A'_n\mathbf{f} + (1 - \frac{1}{c})\mu_n\mathbf{f}\right)\right\}$$

$$= \left\{ (1-\frac{\varepsilon}{2}(1+\frac{1}{c}))(A_n\mathbf{f}-A'_n\mathbf{f}) > \frac{\varepsilon}{2}((1-\frac{1}{c})(A_n\mathbf{f}+A'_n\mathbf{f})+(1-\frac{1}{c})2\alpha)+(1-\frac{1}{c})\varepsilon\mu_n\mathbf{f} \right\}$$

$$\subset \left\{ (A_n\mathbf{f}-A'_n\mathbf{f}) > \eta(2\alpha+A_n\mathbf{f}+A'_n\mathbf{f}) \right\}, \qquad \text{with } 0<\eta := \frac{\varepsilon}{(2-(1+\frac{1}{c})\varepsilon)}(1-\frac{1}{c}), \tag{2.11}$$

and where for the last inclusion we first used $f_k \geq 0$ for all k and then the restriction on c, ε to check that $1-\frac{\varepsilon}{2}(1+\frac{1}{c}) \geq 1-\varepsilon > 0$.

Now we will choose a random $\mathbf{f}^*_{X_{1:n}} \in \mathcal{F}$ (the random choice defining $\mathbf{f}^*_{X_{1:n}}$ will depend on the value of the sample sequence $X_{1:n}$) with the property that

$$\{A_n\mathbf{f}^*_{X_{1:n}} - \mu_n\mathbf{f}^*_{X_{1:n}} > \varepsilon(\alpha + A_n\mathbf{f}^*_{X_{1:n}} + \mu_n\mathbf{f}^*_{X_{1:n}})\} = \bigcup_{\mathbf{f}\in\mathcal{F}} \{A_n\mathbf{f} - \mu_n\mathbf{f} > \varepsilon(\alpha + A_n\mathbf{f} + \mu_n\mathbf{f})\}. \tag{2.12}$$

This can be done as follows: by the pointwise measurability assumption on $\mathcal{F}$, there exist a countable family $\{\mathbf{f}_l\}_{l\in\mathbb{N}} \subset \mathcal{F}$ such that the set at the right–hand side of (2.12) equals

$$\mathcal{E}_1 := \bigcup_{l\in\mathbb{N}} \{A_n\mathbf{f}_l - \mu_n\mathbf{f}_l > \varepsilon(\alpha + A_n\mathbf{f}_l + \mu_n\mathbf{f}_l)\},$$

which is clearly $\sigma(X_{1:n})-$measurable. Then, if $k_{X_{1:n}} : \Omega \to \mathbb{N}$ is the $\sigma(X_{1:n})-$measurable random natural number defined by

$$k_{X_{1:n}} := \mathbf{1}_{\Omega\setminus\mathcal{E}_1} + \min\{l \in \mathbb{N} : A_n\mathbf{f}_l - \mu_n\mathbf{f}_l > \varepsilon(\alpha + A_n\mathbf{f}_l + \mu_n\mathbf{f}_l)\}\mathbf{1}_{\mathcal{E}_1},$$

we can define $\mathbf{f}^*_{X_{1:n}}$ according to the formula

$$\mathbf{f}^*_{X_{1:n}} := \sum_{l\in\mathbb{N}} \mathbf{f}_l \mathbf{1}_{\{k_{X_{1:n}}=l\}},$$

from which the verification of (2.12) follows at once. We remark for the sake of clarity that

$$\mathcal{E}_1 = \{A_n\mathbf{f}^*_{X_{1:n}} - \mu_n\mathbf{f}^*_{X_{1:n}} > \varepsilon(\alpha + A_n\mathbf{f}^*_{X_{1:n}} + \mu_n\mathbf{f}^*_{X_{1:n}})\}. \tag{2.13}$$

Define now $\mathcal{E}_2$ as

$$\mathcal{E}_2 := \{A'_n\mathbf{f}^*_{X_{1:n}} - \mu_n\mathbf{f}^*_{X_{1:n}} \leq \frac{\varepsilon}{c}(\alpha + A'_n\mathbf{f}^*_{X_{1:n}} + \mu_n\mathbf{f}^*_{X_{1:n}})\}. \tag{2.14}$$

Then, by the argument leading to (2.11) and an obvious inclusion we get that

$$\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \leq \mathbb{P}\left(\bigcup_{\mathbf{f}\in\mathcal{F}} \left\{ A_n\mathbf{f} - A'_n\mathbf{f} > \eta(2\alpha + A_n\mathbf{f} + A'_n\mathbf{f}) \right\} \right). \tag{2.15}$$

Notice that, since the functions in $\mathcal{F}$ are nonnegative, the inclusion

$$\Omega \setminus \mathcal{E}_2 \subset \{A'_n\mathbf{f}^*_{X_{1:n}} - \mu_n\mathbf{f}^*_{X_{1:n}} > \frac{\varepsilon}{c}(\alpha + \mu_n\mathbf{f}^*_{X_{1:n}})\}$$

holds. Using the $\sigma(X_{1:n})-$measurability of $\mathbf{1}_{\mathcal{E}_1}$, Lemma 2.5 with $u = +1$, and taking

$$n \geq \frac{c^2\gamma}{4\alpha(\alpha+1)\varepsilon^2} \tag{2.16}$$

we arrive at

$$\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) = \mathbb{E}\left[\mathbf{1}_{\mathcal{E}_1}(1 - \mathbb{P}_n(\Omega \setminus \mathcal{E}_2))\right] \geq \left(1 - \frac{c^2}{4\alpha(\alpha+1)\varepsilon^2 n}\right)\mathbb{P}(\mathcal{E}_1) \geq (1-\frac{1}{\gamma})\mathbb{P}(\mathcal{E}_1). \tag{2.17}$$

This leads, by the definition of $\mathcal{E}_1$ and (2.15), to the following conclusion:

Lemma 2.6. *Under the hypotheses of Theorem 2.2 (with $B = 1$), for every n as in (2.16) and for η as in (2.11), we have*

$$\mathbb{P}\left(\bigcup_{\mathbf{f}\in\mathcal{F}}\left\{A_n\mathbf{f} - \mu_n\mathbf{f} > \varepsilon(\alpha + A_n\mathbf{f} + \mu_n\mathbf{f})\right\}\right) \le \frac{\gamma}{\gamma-1}\mathbb{P}\left(\bigcup_{\mathbf{f}\in\mathcal{F}}\left\{A_n\mathbf{f} - A_n'\mathbf{f} > \eta(2\alpha + A_n\mathbf{f} + A_n'\mathbf{f})\right\}\right). \tag{2.18}$$

▷ *Introduction of Random Signs.* To estimate the probability at the right–hand side of (2.18) we use the following observation: if $\mathbf{U} = (U_1, \dots, U_n)$ is a sequence of independent Rademacher random variables, independent of $(X_k)_k$ and $(X_k')_k$, and $\mathbf{f}_1, \dots, \mathbf{f}_L \in \mathcal{F}$ are given ($\mathbf{f}_j = (f_{j,k})_k$), then the joint distributions of

$$(U_k f_{j,k}(X_k) - U_k f_{j,k}(X_k'), f_{j,k}(X_k) + f_{j,k}(X_k'))_{j,k}$$

and

$$(f_{j,k}(X_k) - f_{j,k}(X_k'), f_{j,k}(X_k) + f_{j,k}(X_k'))_{j,k}$$

are the same for fixed values of $U_1, \dots, U_n$ (they correspond to interchanging X_k and X_k' for *some* k). Therefore we have that (remember the notation $\mathbf{Uf} := (U_k f_k)$)

$$\mathbb{P}\left(\bigcup_{\mathbf{f}\in\mathcal{F}}\{A_n\mathbf{Uf} - A_n'\mathbf{Uf} > \eta(2\alpha + A_n\mathbf{f} + A_n'\mathbf{f})\} \mid U_{1:n}\right) = \mathbb{P}\left(\bigcup_{\mathbf{f}\in\mathcal{F}}\{A_n\mathbf{f} - A_n'\mathbf{f} > \eta(2\alpha + A_n\mathbf{f} + A_n'\mathbf{f})\}\right),$$

$\mathbb{P}-$a.s. Owing to the random exchange of signs, we finally get

$$\begin{aligned}&\mathbb{P}\left(\bigcup_{\mathbf{f}\in\mathcal{F}}\{A_n\mathbf{f} - A_n'\mathbf{f} > \eta(2\alpha + A_n\mathbf{f} + A_n'\mathbf{f})\}\right)\\ &= \mathbb{P}\left(\bigcup_{\mathbf{f}\in\mathcal{F}}\{A_n\mathbf{Uf} - A_n'\mathbf{Uf} > \eta(2\alpha + A_n\mathbf{f} + A_n'\mathbf{f})\}\right) \le 2\mathbb{P}\left(\bigcup_{\mathbf{f}\in\mathcal{F}}\{A_n\mathbf{Uf} > \eta(\alpha + A_n\mathbf{f})\}\right),\end{aligned} \tag{2.19}$$

where the last inequality follows from

$$\{A_n\mathbf{Uf} - A_n'\mathbf{Uf} > \eta(2\alpha + A_n\mathbf{f} + A_n'\mathbf{f})\} \subset \{A_n\mathbf{Uf} > \eta(\alpha + A_n\mathbf{f})\} \cup \{A_n'\mathbf{Uf} < -\eta(\alpha + A_n'\mathbf{f})\}$$

and from the symmetry of $(U_k)_k$. We have thus proved the following

Lemma 2.7. *Under the hypothesis of Theorem 2.2 (with $B = 1$) and the conventions in Section 2.2.1, for any $n \in \mathbb{N}$, any $\eta > 0$, and any sequence $\mathbf{U} = (U_1, \dots, U_n)$ of independent Rademacher random variables independent of $X_{1:n}$ and $X_{1:n}'$, we have*

$$\mathbb{P}\left(\bigcup_{\mathbf{f}\in\mathcal{F}}\left\{A_n\mathbf{f} - A_n'\mathbf{f} > \eta(2\alpha + A_n\mathbf{f} + A_n'\mathbf{f})\right\}\right) \le 2\mathbb{P}\left(\bigcup_{\mathbf{f}\in\mathcal{F}}\{A_n\mathbf{Uf} > \eta(\alpha + A_n\mathbf{f})\}\right). \tag{2.20}$$

▷ *Introduction of Covering Numbers.* Now we bound the probability at the right-hand side of (2.20) by covering $\mathcal{F}$ with random balls of radius $\delta > 0$ (with respect to the empirical L_1 norm at $X_{1:n}$) and we estimate the deviation for each ball center by using Hoeffding's inequality.

Given $\delta > 0$ and $\omega \in \Omega$ there exists $N(\omega) := \mathcal{N}_1(\delta, \mathcal{F}, X_{1:n}(\omega))$ sequences of functions $\mathbf{f}_{X_{1:n}(\omega),1}, \dots, \mathbf{f}_{X_{1:n}(\omega),N(\omega)}$ (we stress the dependence on $(X_k(\omega))_{k=1}^n$) satisfying the following property: for every $\mathbf{f} \in \mathcal{F}$ and some $j(\mathbf{f}) \in \{1, \dots, N(\omega)\}$

$$\frac{1}{n}\sum_{k=1}^{n} |f_k(X_k(\omega)) - f_{X_{1:n}(\omega),j(\mathbf{f}),k}(X_k(\omega))| \le \delta.$$

Without loss of generality (truncate if necessary), we can assume that each $f_{X_{1:n},j,k}$ takes values in $[0,1]$. Triangular inequalities yield

$$A_n \mathbf{U}\mathbf{f}_{X_{1:n},j(\mathbf{f})} + \delta \ge A_n \mathbf{U}\mathbf{f}, \qquad A_n \mathbf{f} \ge A_n \mathbf{f}_{X_{1:n},j(\mathbf{f})} - \delta,$$

which leads to the inclusion (available for any $\eta > 0$ and in particular for η given in (2.11))

$$\bigcup_{\mathbf{f}\in\mathcal{F}} \{A_n \mathbf{U}\mathbf{f} > \eta(\alpha + A_n \mathbf{f})\} \subset \bigcup_{j\in\{1,\dots,\mathcal{N}_1(\delta,\mathcal{F},X_{1:n})\}} \{\delta + A_n \mathbf{U}\mathbf{f}_{X_{1:n},j} > \eta(\alpha - \delta + A_n \mathbf{f}_{X_{1:n},j})\}.$$

As a consequence, we arrive to the following estimate:

Lemma 2.8. *Under the hypothesis of Theorem 2.2 (with $B = 1$) and the conventions in Section 2.2.1 (see also Section 1.4), for any $n \in \mathbb{N}$, any $\eta > 0$, any sequence $\mathbf{U} = (U_1, \dots, U_n)$ of independent Rademacher random variables independent of $X_{1:n}$ and $X'_{1:n}$, any $\delta > 0$, and any (minimal) cover $\mathbf{f}_{X_{1:n},1}, \dots, \mathbf{f}_{X_{1:n},\mathcal{N}_1(\delta,\mathcal{F},X_{1:n})}$ of $\mathcal{F}$ with respect to the empirical L_1 norm at $X_{1:n}$*

$$\mathbb{P}_n\left(\bigcup_{\mathbf{f}\in\mathcal{F}} \{A_n \mathbf{U}\mathbf{f} > \eta(\alpha + A_n \mathbf{f})\}\right) \le \mathcal{N}_1(\delta, \mathcal{F}, X_{1:n}) \max_j \mathbb{P}_n\left(\delta + A_n \mathbf{U}\mathbf{f}_{X_{1:n},j} > \eta(\alpha - \delta + A_n \mathbf{f}_{X_{1:n},j})\right). \tag{2.21}$$

▷ *Bounding uniformly the tails in (2.21), for an appropriate δ, via Hoeffding's Inequality.* We introduce at this step the parameter $\gamma' > 1$ by setting

$$\delta = \frac{\eta}{\gamma'(\eta+1)}\alpha. \tag{2.22}$$

This choice of δ allows us to rewrite

$$\mathbb{P}_n\left(\delta + A_n \mathbf{U}\mathbf{f}_{X_{1:n},j} > \eta(\alpha - \delta + A_n \mathbf{f}_{X_{1:n},j})\right) = \mathbb{P}_n\left(A_n \mathbf{U}\mathbf{f}_{X_{1:n},j} > \eta((1 - \frac{1}{\gamma'})\alpha + A_n \mathbf{f}_{X_{1:n},j})\right).$$

Assume that $A_n \mathbf{f}_{X_{1:n},j} \neq 0$ (and thus > 0), otherwise the above probability is zero. Apply the Hoeffding inequality [GKKW02, Lemma A.3] (restricted to the right tail) together with the inequality $-f_k \le U_k f_k \le f_k$ to get

$$\begin{aligned}
\mathbb{P}_n\left(A_n \mathbf{U}\mathbf{f}_{X_{1:n},j} > \eta((1 - \frac{1}{\gamma'})\alpha + A_n \mathbf{f}_{X_{1:n},j})\right) &\le \exp\left(-\frac{\eta^2((1-\frac{1}{\gamma'})\alpha + A_n \mathbf{f}_{X_{1:n},j})^2}{2A_n \mathbf{f}^2_{X_{1:n},j}} n\right) \\
&\le \exp\left(-\frac{\eta^2((1-\frac{1}{\gamma'})\alpha + A_n \mathbf{f}_{X_{1:n},j})^2}{2A_n \mathbf{f}_{X_{1:n},j}} n\right) \\
&\le \exp\left(-2\eta^2 n(1 - \frac{1}{\gamma'})\alpha\right),
\end{aligned} \tag{2.23}$$

where at the second inequality, we used that $0 \leq f_{X_{1:n},j,k} \leq 1$ for all k, and at the last inequality we used the minimization $\frac{(a+y)^2}{y} \geq 4a$ valid for $(a,y) \in (0,\infty) \times (0,\infty)$.
$\triangleright$ *Conclusion for* $B = 1$. A combination of Lemma 2.8 together with the choice (2.22) of δ and the estimate (2.23) gives rise to the following result:

Lemma 2.9. *Under the hypothesis of Theorem 2.2 (with* $B = 1$*) and the conventions in Section 2.2.1 (see also Section 1.4), for any* $n \in \mathbb{N}$*, any* $\eta > 0$ *and any sequence* $\mathbf{U} = (U_1, \ldots, U_n)$ *of independent Rademacher random variables independent of* $X_{1:n}$*, we have*

$$\mathbb{P}_n\left(\bigcup_{\mathbf{f}\in\mathcal{F}}\{A_n\mathbf{U}\mathbf{f} > \eta(\alpha + A_n\mathbf{f})\}\right) \leq \mathcal{N}_1(\frac{\alpha}{\gamma'}\frac{\eta}{\eta+1}, \mathcal{F}, X_{1:n}) \times \exp\left(-2\eta^2 n(1-\frac{1}{\gamma'})\alpha\right). \tag{2.24}$$

We conclude thanks to Lemmas 2.6, 2.7, and 2.9, via the substitution $\eta = \frac{\varepsilon}{(2-(1+\frac{1}{c})\varepsilon)}(1-\frac{1}{c})$ (see (2.11)), that the left hand side of (2.2) is bounded by

$$\frac{2\gamma}{\gamma-1}\mathbb{E}\left[\mathcal{N}_1(\frac{\alpha}{\gamma'}\frac{(c-1)}{(c-\varepsilon)}\frac{\varepsilon}{2}, \mathcal{F}, X_{1:n})\right]\exp\left(-\frac{2\varepsilon^2}{(2-(1+\frac{1}{c})\varepsilon)^2}(1-\frac{1}{c})^2 n(1-\frac{1}{\gamma'})\alpha\right),$$

provided that n satisfies (2.16). This gives the conclusion for the case $B = 1$.
$\triangleright$ *The conclusion for arbitrary* $B > 0$. It comes again from homogenization: if we consider

$$\frac{\mathcal{F}}{B} := \{\frac{1}{B}\mathbf{f} : \mathbf{f} \in \mathcal{F}\}, \tag{2.25}$$

then the case $B = 1$ together with the equality

$$\mathcal{N}_1(\delta, \frac{\mathcal{F}}{B}, X_{1:n}) = \mathcal{N}_1(\delta B, \mathcal{F}, X_{1:n}), \tag{2.26}$$

imply (2.2) (divide by B inside the sets at the left hand side of (2.2)) . □

Proof of Inequality (2.3)

The proof is basically identical to that of Inequality (2.2), thus we will only indicate the adjustments. We deal only with $B = 1$, the case $B > 0$ follows as before.

Similarly to (2.12), let $\mathbf{f}^{**}_{X_{1:n}} \in \mathcal{F}$ be such that

$$\mathcal{E}'_1 := \bigcup_{f\in\mathcal{F}}\{\mu_n\mathbf{f} - A_n\mathbf{f} > \varepsilon(\alpha + \mu_n\mathbf{f} + A_n\mathbf{f})\} = \{\mu_n\mathbf{f}^{**}_{X_{1:n}} - A_n\mathbf{f}^{**}_{X_{1:n}} > \varepsilon(\alpha + \mu_n\mathbf{f}^{**}_{X_{1:n}} + A_n\mathbf{f}^{**}_{X_{1:n}})\},$$

and set

$$\mathcal{E}'_2 := \{\mu_n\mathbf{f}^{**}_{X_{1:n}} - A'_n\mathbf{f}^{**}_{X_{1:n}} \leq \frac{\varepsilon}{c}(\alpha + A'_n\mathbf{f}^{**}_{X_{1:n}} + \mu_n\mathbf{f}^{**}_{X_{1:n}})\}.$$

By choosing n large enough (as in (2.16)) we get as for (2.17) by using this time Lemma 2.5 with $u = -1$

$$\mathbb{P}\left(\mathcal{E}'_1\right) \leq \frac{\gamma}{\gamma-1}\mathbb{P}\left(\mathcal{E}'_1 \cap \mathcal{E}'_2\right).$$

Moreover, simple computations as before yield

$$\mathbb{P}\left(\bigcup_{f\in\mathcal{F}}\{\mu_n\mathbf{f} - A_n\mathbf{f} > \varepsilon(\alpha + \mu_n\mathbf{f} + A_n\mathbf{f})\} \cap \{\mu_n\mathbf{f} - A'_n\mathbf{f} \leq \frac{\varepsilon}{c}(\alpha + \mu_n\mathbf{f} + A'_n\mathbf{f})\}\right)$$

$$\leq \mathbb{P}\left(\bigcup_{f\in\mathcal{F}}\{(1+\frac{\varepsilon}{2}(1+\frac{1}{c}))(A'_n\mathbf{f}-A_n\mathbf{f})>\frac{\varepsilon}{2}(1-\frac{1}{c})(2\alpha+A'_n\mathbf{f}+A_n\mathbf{f})+\varepsilon(1-\frac{1}{c})\mu_n\mathbf{f}\}\right)$$
$$\leq \mathbb{P}\left(\bigcup_{\mathbf{f}\in\mathcal{F}}\left\{A'_n\mathbf{f}-A_n\mathbf{f}>\eta'(2\alpha+A_n\mathbf{f}+A'_n\mathbf{f})\right\}\right),$$

where

$$\eta' := \frac{\frac{\varepsilon}{2}(1-\frac{1}{c})}{1+\frac{\varepsilon}{2}(1+\frac{1}{c})}. \tag{2.27}$$

Gathering the arguments, we deduce again that the probability to bound is such that

$$\mathbb{P}\left(\mathcal{E}'_1\right) \leq \frac{\gamma}{\gamma-1}\mathbb{P}\left(\bigcup_{\mathbf{f}\in\mathcal{F}}\left\{A'_n\mathbf{f}-A_n\mathbf{f}>\eta'(2\alpha+A_n\mathbf{f}+A'_n\mathbf{f})\right\}\right).$$

This is similar to (2.18).

From here, we follow exactly the same arguments used before to arrive to the inequality that corresponds to the deduction of (2.24) obtaining this time, for η' given by (2.27), that the left hand side of (2.3) is bounded by

$$\frac{2\gamma}{\gamma-1}\mathbb{E}\left[\mathcal{N}_1(\frac{\alpha}{\gamma'}\frac{(c-1)}{(c+c\varepsilon)}\frac{\varepsilon}{2},\mathcal{F},X_{1:n})\right]\times\exp\left(-\frac{2\varepsilon^2}{(2+(1+\frac{1}{c})\varepsilon)^2}(1-\frac{1}{c})^2n(1-\frac{1}{\gamma'})\alpha\right).$$

This gives the desired conclusion. □

2.3 Proof of Theorem 2.3

We will use the notation

$$Z_k := (X_k, Y_k), \qquad Z'_k := (X'_k, Y'_k). \tag{2.28}$$

As before, we will start assuming that $B = 1$, and we will give the proofs in a series of partial results as in the proof of (2.2) .

▷ *First Symmetrization.* Fix $u = \pm 1$ and recall the notation (2.9). We have to bound

$$\mathbb{P}\left(\bigcup_{f\in\mathcal{F}}\{u(\widetilde{\mu_n}\mathbf{g}_f-\widetilde{A}_n\mathbf{g}_f)\geq\varepsilon(\alpha+\widetilde{\mu_n}\mathbf{g}_f)\}\right). \tag{2.29}$$

To begin with, we will prove:

Lemma 2.10. *With the notation in Section 2.2.1, and under the hypothesis of Theorem 2.3 (with $B = 1$)*

$$\mathbb{P}\left(\bigcup_{f\in\mathcal{F}}\{u(\widetilde{\mu_n}\mathbf{g}_f-\widetilde{A}_n\mathbf{g}_f)\geq\varepsilon(\alpha+\widetilde{\mu_n}\mathbf{g}_f)\}\right)\leq\frac{\gamma}{\gamma-1}\mathbb{P}\left(\bigcup_{f\in\mathcal{F}}\{u(\widetilde{A}'_n\mathbf{g}_f-\widetilde{A}_n\mathbf{g}_f)\geq(\frac{\gamma'-1}{\gamma'})\varepsilon(\alpha+\widetilde{\mu_n}\mathbf{g}_f)\}\right), \tag{2.30}$$

provided that

$$n\geq\frac{\gamma}{4\alpha(\alpha+1)}\left(\frac{\gamma'}{\varepsilon}\right)^2. \tag{2.31}$$

Indeed, let us choose, as in (2.12), a random function $f_n := f_{n,Z_{1:n}} \in \mathcal{F}$ in such a way that

$$\{u(\widetilde{\mu_n}\mathbf{g}_{f_n} - \widetilde{A}_n\mathbf{g}_{f_n}) \geq \varepsilon(\alpha + \widetilde{\mu_n}\mathbf{g}_{f_n})\} = \bigcup_{f\in\mathcal{F}} \{u(\widetilde{\mu_n}\mathbf{g}_f - \widetilde{A}_n\mathbf{g}_f) \geq \varepsilon(\alpha + \widetilde{\mu_n}\mathbf{g}_f)\}. \tag{2.32}$$

Then, for $\gamma' > 1$,

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{f\in\mathcal{F}} \{u(\widetilde{A}'_n\mathbf{g}_f - \widetilde{A}_n\mathbf{g}_f) \geq (1-\frac{1}{\gamma'})\varepsilon(\alpha + \widetilde{\mu_n}\mathbf{g}_f)\}\right) \\
&\geq \mathbb{P}\left(\{u(\widetilde{\mu_n}\mathbf{g}_{f_n} - \widetilde{A}_n\mathbf{g}_{f_n}) \geq \varepsilon(\alpha + \widetilde{\mu_n}\mathbf{g}_{f_n})\} \cap \{u(\widetilde{\mu_n}\mathbf{g}_{f_n} - \widetilde{A}'_n\mathbf{g}_{f_n}) \leq \frac{\varepsilon}{\gamma'}(\alpha + \widetilde{\mu_n}\mathbf{g}_f)\}\right) \\
&=: \mathbb{P}\left(\mathcal{E}_1 \cap \mathcal{E}_2\right) = \mathbb{E}\left[\mathbf{1}_{\mathcal{E}_1}(1 - \tilde{\mathbb{P}}_n(\Omega \setminus \mathcal{E}_2))\right],
\end{aligned} \tag{2.33}$$

where this time

$$\begin{aligned}
\mathcal{E}_1 &:= \{u(\widetilde{\mu_n}\mathbf{g}_{f_n} - \widetilde{A}_n\mathbf{g}_{f_n}) \geq \varepsilon(\alpha + \widetilde{\mu_n}\mathbf{g}_{f_n})\}, \\
\mathcal{E}_2 &:= \{u(\widetilde{\mu_n}\mathbf{g}_{f_n} - \widetilde{A}'_n\mathbf{g}_{f_n}) \leq \frac{\varepsilon}{\gamma'}(\alpha + \widetilde{\mu_n}\mathbf{g}_f)\}
\end{aligned}$$

(compare with (2.13), (2.14), and the arguments that follow). Now notice that, in virtue of (2.5) and the assumption $B = 1$, we have

$$|g_f^k(X_k, Y_k)| = |(f(X_k) - \Phi_k(X_k))(f(X_k) - 2Y_k + \Phi_k(X_k))| \leq |f(X_k) - \Phi_k(X_k)|,$$

which leads to the (crucial) inequality

$$\widetilde{\mu_n}\mathbf{g}_f^2 \leq \mu_n|\mathbf{f} - \mathbf{\Phi}_{1:n}|^2 = \widetilde{\mu}_n(|\mathbf{f} - \mathbf{y}_{1:n}|^2 - |\mathbf{y}_{1:n} - \mathbf{\Phi}_{1:n}|^2) = \widetilde{\mu_n}\mathbf{g}_f, \tag{2.34}$$

where $\mathbf{f} = (f, \ldots, f)$ as explained in introduction. Now we estimate: proceeding as in (2.10) and using (2.34)

$$\begin{aligned}
\tilde{\mathbb{P}}_n(\Omega \setminus \mathcal{E}_2) &= \tilde{\mathbb{P}}_n(u(\widetilde{\mu_n}\mathbf{g}_{f_n} - \widetilde{A}'_n\mathbf{g}_{f_n}) > \frac{\varepsilon}{\gamma'}(\alpha + \widetilde{\mu_n}\mathbf{g}_{f_n})) \\
&\leq \frac{\gamma'^2}{n\varepsilon^2}\frac{\widetilde{\mu_n}\mathbf{g}_{f_n}^2 - (\widetilde{\mu_n}\mathbf{g}_{f_n})^2}{(\alpha + \widetilde{\mu_n}\mathbf{g}_{f_n})^2} \leq \frac{\gamma'^2}{n\varepsilon^2}\frac{\widetilde{\mu_n}\mathbf{g}_{f_n}(1 - \widetilde{\mu_n}\mathbf{g}_{f_n})}{(\alpha + \widetilde{\mu_n}\mathbf{g}_{f_n})^2} \leq \frac{\gamma'^2}{4n\alpha(1+\alpha)\varepsilon^2},
\end{aligned} \tag{2.35}$$

where we used the same optimization argument as after (2.10). Together, (2.33) and (2.35) lead to the following conclusion: if n satisfies (2.31), then

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{f\in\mathcal{F}} \{u(\widetilde{A}'_n\mathbf{g}_f - \widetilde{A}_n\mathbf{g}_f) \geq (1-\frac{1}{\gamma'})\varepsilon(\alpha + \widetilde{\mu_n}\mathbf{g}_f)\}\right) \\
&\quad \geq (1 - \frac{1}{\gamma})\mathbb{P}\left(u(\widetilde{\mu_n}\mathbf{g}_{f_n} - \widetilde{A}_n\mathbf{g}_{f_n}) \geq \varepsilon(\alpha + \widetilde{\mu_n}\mathbf{g}_{f_n})\right),
\end{aligned} \tag{2.36}$$

which implies at once (2.30) by the choice (2.32) of f_n. This proves Lemma 2.10.

▷ *Application of Theorem 2.2.* In this step we will bound the probability at the left-hand side of (2.36) using the decomposition in the following lemma.

Lemma 2.11. *With the notation in Section 2.2.1, and under the hypothesis of Theorem 2.3 (with $B = 1$) consider, for every $f \in \mathcal{F}$, the events*

$$\begin{aligned}
\mathcal{E}_f &:= \{\widetilde{A}'_n \mathbf{g}_f - \widetilde{A}_n \mathbf{g}_f \geq (1 - \frac{1}{\gamma'})\varepsilon(\alpha + \widetilde{\mu_n}\mathbf{g}_f)\},\\
\mathcal{E}_{f,1} &:= \{\widetilde{A}_n \mathbf{g}_f^2 - \widetilde{\mu_n}\mathbf{g}_f^2 \leq \varepsilon(\alpha + \widetilde{A}_n \mathbf{g}_f^2 + \widetilde{\mu_n}\mathbf{g}_f^2)\},\\
\mathcal{E}'_{f,1} &:= \{\widetilde{A}'_n \mathbf{g}_f^2 - \widetilde{\mu_n}\mathbf{g}_f^2 \leq \varepsilon(\alpha + \widetilde{A}'_n \mathbf{g}_f^2 + \widetilde{\mu_n}\mathbf{g}_f^2)\}.
\end{aligned} \tag{2.37}$$

Then for $u = \pm 1$ fixed,

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{f\in\mathcal{F}} \{u(\widetilde{A}'_n \mathbf{g}_f - \widetilde{A}_n \mathbf{g}_f) \geq (1 - \frac{1}{\gamma'})\varepsilon(\alpha + \widetilde{\mu_n}\mathbf{g}_f)\}\right) \leq \mathbb{P}\left(\bigcup_{f\in\mathcal{F}} \mathcal{E}_f \cap \mathcal{E}_{f,1} \cap \mathcal{E}'_{f,1}\right)\\
&+ 4\frac{\gamma}{\gamma - 1}\mathbb{E}\left[\mathcal{N}_1(\frac{\alpha}{\gamma'}\frac{(c-1)}{(c-\varepsilon)}\frac{\varepsilon}{2}, \{\mathbf{g}_f^2 : f \in \mathcal{F}\}, Z_{1:n})\right] \times \exp\left(-\frac{2\varepsilon^2}{(2 - (1 + \frac{1}{c})\varepsilon)^2}(1 - \frac{1}{c})^2(1 - \frac{1}{\gamma'})\alpha n\right)
\end{aligned} \tag{2.38}$$

provided that

$$n \geq \frac{\gamma}{4\alpha(\alpha + 1)}\left(\frac{c}{\varepsilon}\right)^2. \tag{2.39}$$

Indeed, note first that we can assume that $u = 1$, because all the arguments hold exchanging the roles of $X_{1:n}$ and $X'_{1:n}$. Then from

$$\mathcal{E}_f \subset (\mathcal{E}_f \cap \mathcal{E}_{f,1} \cap \mathcal{E}'_{f,1}) \cup (\Omega \setminus \mathcal{E}_{f,1}) \cup (\Omega \setminus \mathcal{E}'_{f,1}),$$

we get, via the union bound and the symmetry of $\mathcal{E}_{f,1}, \mathcal{E}'_{f,1}$, that

$$\mathbb{P}\left(\bigcup_{f\in\mathcal{F}} \mathcal{E}_f\right) \leq \mathbb{P}\left(\bigcup_{f\in\mathcal{F}} \mathcal{E}_f \cap \mathcal{E}_{f,1} \cap \mathcal{E}'_{f,1}\right) + 2\mathbb{P}\left(\bigcup_{f\in\mathcal{F}} \{\widetilde{A}_n \mathbf{g}_f^2 - \widetilde{\mu_n}\mathbf{g}_f^2 > \varepsilon(\alpha + \widetilde{A}_n \mathbf{g}_f^2 + \widetilde{\mu_n}\mathbf{g}_f^2)\}\right). \tag{2.40}$$

Now, by the assumption $B = 1$, we can assume that

$$\sup_{z\in\mathbb{R}^d\times\mathbb{R}} (g_f^k)^2(z) \leq 1 \tag{2.41}$$

for all $k \in \{1, \dots, n\}$ (indeed notice that, for fixed k, such inequality holds for $\mathbb{P}-$a.e. z with respect to the law of (X_k, Y_k) in virtue of (2.5) and the definition (2.8) of g_f^k).

With this, (2.38) follows by an estimation, via (2.2), of the second probability at the right hand side of (2.40), noticing that such estimate holds if n satisfies (2.39). This proves Lemma 2.11.

▷ *Second Symmetrization and Introduction of Random Signs.* So far we have proved, by a combination of Lemmas 2.10 and 2.11, that if n satisfies and (2.31) and (2.39), i.e., if (2.7) holds (for $B = 1$), then (2.29) is bounded by $\frac{\gamma}{\gamma-1}$ times the right–hand side of (2.38). Our goal now is to have a bound similar for the probability

$$\mathbb{P}\left(\bigcup_{f\in\mathcal{F}} \mathcal{E}_f \cap \mathcal{E}_{f,1} \cap \mathcal{E}'_{f,1}\right).$$

To do so, we notice that, by (2.34)

$$\begin{aligned}\mathcal{E}_{f,1}\cap\mathcal{E}'_{f,1} &= \{(1+\varepsilon)\widetilde{\mu_n}\mathbf{g}_f^2 \geq (1-\varepsilon)\widetilde{A}_n\mathbf{g}_f^2 - \varepsilon\alpha\}\cap\{(1+\varepsilon)\widetilde{\mu_n}\mathbf{g}_f^2 \geq (1-\varepsilon)\widetilde{A}'_n\mathbf{g}_f^2 - \varepsilon\alpha\}\\ &\subset \{(1+\varepsilon)\widetilde{\mu_n}\mathbf{g}_f \geq (1-\varepsilon)\widetilde{A}_n\mathbf{g}_f^2 - \varepsilon\alpha\}\cap\{(1+\varepsilon)\widetilde{\mu_n}\mathbf{g}_f \geq (1-\varepsilon)\widetilde{A}'_n\mathbf{g}_f^2 - \varepsilon\alpha\}\end{aligned}$$

and therefore,

$$\begin{aligned}&\mathcal{E}_f\cap\mathcal{E}_{f,1}\cap\mathcal{E}'_{f,1}\\ &= \{(1+\varepsilon)(\widetilde{A}'_n\mathbf{g}_f - \widetilde{A}_n\mathbf{g}_f) \geq (1+\varepsilon)(1-\frac{1}{\gamma'})\varepsilon(\alpha+\widetilde{\mu_n}\mathbf{g}_f)\}\cap\mathcal{E}_{f,1}\cap\mathcal{E}'_{f,1}\\ &\subset \{(1+\varepsilon)(\widetilde{A}'_n\mathbf{g}_f - \widetilde{A}_n\mathbf{g}_f) \geq \frac{1}{2}(1-\frac{1}{\gamma'})\varepsilon((1+\varepsilon)\alpha + (1-\varepsilon)\widetilde{A}_n\mathbf{g}_f^2 - \varepsilon\alpha)\\ &\qquad\qquad + \frac{1}{2}(1-\frac{1}{\gamma'})\varepsilon((1+\varepsilon)\alpha + (1-\varepsilon)\widetilde{A}'_n\mathbf{g}_f^2 - \varepsilon\alpha)\}\\ &= \{(1+\varepsilon)(\widetilde{A}'_n\mathbf{g}_f - \widetilde{A}_n\mathbf{g}_f) \geq \frac{1}{2}(1-\frac{1}{\gamma'})\varepsilon(\alpha+(1-\varepsilon)\widetilde{A}_n\mathbf{g}_f^2) + \frac{1}{2}(1-\frac{1}{\gamma'})\varepsilon(\alpha+(1-\varepsilon)\widetilde{A}'_n\mathbf{g}_f^2)\}. \quad (2.42)\end{aligned}$$

This last event has the form

$$\{\widetilde{A}_n\mathbf{g}_f - \widetilde{A}'_n\mathbf{g}_f \geq (c_1 + c_2\widetilde{A}_n\mathbf{g}_f^2) + (c_1 + c_2\widetilde{A}'_n\mathbf{g}_f^2)\},$$

with

$$c_0 = \frac{1}{2}(1-\frac{1}{\gamma'})\frac{1}{(1+\varepsilon)}\varepsilon, \quad c_1 = c_0\alpha, \quad c_2 = c_0(1-\varepsilon), \tag{2.43}$$

which by random exchange of signs (see for instance the arguments leading to (2.19)) allows us to conclude that for any sequence $\mathbf{U} = (U_k)_k$ of independent Rademacher random variables independent from $(Z_k)_k$ and $(Z'_k)_k$,

$$\begin{aligned}&\mathbb{P}\left(\bigcup_{f\in\mathcal{F}}\{\widetilde{A}_n\mathbf{U}\mathbf{g}_f - \widetilde{A}'_n\mathbf{U}\mathbf{g}_f \geq (c_1 + c_2\widetilde{A}_n\mathbf{g}_f^2) + (c_1 + c_2\widetilde{A}'_n\mathbf{g}_f^2)\}\right)\\ &= \mathbb{P}\left(\bigcup_{f\in\mathcal{F}}\{\widetilde{A}_n\mathbf{g}_f - \widetilde{A}'_n\mathbf{g}_f \geq (c_1 + c_2\widetilde{A}_n\mathbf{g}_f^2) + (c_1 + c_2\widetilde{A}'_n\mathbf{g}_f^2)\}\right).\end{aligned}$$

This observation together with (2.42), the triangle inequality, the union bound, and symmetry (note that $(Z_{1:n}, U_{1:n})$ and $(Z'_{1:n}, -U_{1:n})$ have the same distribution), allows us to conclude the following

Lemma 2.12. *Under the hypothesis of Theorem 2.3 (with $B = 1$), with the notations from Section 2.2.1 and in (2.37), for any $n \in \mathbb{N}$, and for any sequence $\mathbf{U} = (U_1, \dots, U_n)$ of independent Rademacher random variables independent of $(X_k, Y_k)_{k\in\{1,\dots n\}}$ and $(X'_k, Y'_k)_{k\in\{1,\dots n\}}$, and for c_1, c_2, as in (2.43), the estimate*

$$\mathbb{P}\left(\bigcup_{f\in\mathcal{F}}(\mathcal{E}_f\cap\mathcal{E}_{f,1}\cap\mathcal{E}'_{f,1})\right) \leq 2\mathbb{P}\left(\bigcup_{f\in\mathcal{F}}\{\widetilde{A}_n\mathbf{U}\mathbf{g}_f \geq c_1 + c_2\widetilde{A}_n\mathbf{g}_f^2\}\right) \tag{2.44}$$

holds.

▷ *Further Introduction of Covering Numbers.* Following ideas already used we will estimate the probability at the right hand side of (2.44) by first conditioning with respect to the data $Z_{1:n}$ (see (2.28)) and then introducing covering numbers in order to reduce the estimates to finitely many applications of a Hoeffding-type inequality.

For $z_1, \dots, z_n \in \mathbb{R}^d \times \mathbb{R}$ given, consider an event of the form

$$\bigcup_{f \in \mathcal{F}} \{\frac{1}{n}\sum_{k=1}^{n} U_k g_f^k(z_k) \geq c_1 + c_2 \frac{1}{n}\sum_{k=1}^{n} (g_f^k)^2(z_k)\}, \tag{2.45}$$

(like the one in (2.44) for fixed values $Z_{1:n} = z_{1:n}$). For any $\delta > 0$, there exists $N(z_{1:n}) := \mathcal{N}_1(\delta, \{\mathbf{g}_f : f \in \mathcal{F}\}, z_{1:n})$ sequences of functions $\{\mathbf{g}_{z_{1:n},1}, \dots, \mathbf{g}_{z_{1:n},N(z_{1:n})}\}$ (each $\mathbf{g}$ has the form $\mathbf{g} = (g^k)_k$) with the property that for every $f \in \mathcal{F}$

$$\frac{1}{n}\sum_{k=1}^{n} |g_{z_{1:n},j_f}^k(z_k) - g_f^k(z_k)| \leq \delta,$$

for some $j_f \in \{1, \dots, N(z_{1:n})\}$. Without loss of generality (by (2.41)) we can assume that

$$\sup_{z \in \mathbb{R}^d \times \mathbb{R}} |g_j^k(z)| \leq 1$$

for $(j,k) \in \{1, \dots, N(z_{1:n})\} \times \{1, \dots, n\}$. We get, via the triangle inequality and the inequality

$$|(g_{z_{1:n},j_f}^k)^2(z_k) - (g_f^k)^2(z_k)| \leq |g_{z_{1:n},j_f}^k(z_k) + g_f^k(z_k)||g_{z_{1:n},j_f}^k(z_k) - g_f^k(z_k)| \leq 2|g_{z_{1:n},j_f}^k(z_k) - g_f^k(z_k)|$$

that for every $f \in \mathcal{F}$

$$\delta + \frac{1}{n}\sum_{k=1}^{n} U_k g_{z_{1:n},j_f}^k(z_k) \geq \frac{1}{n}\sum_{k=1}^{n} U_k g_f^k(z_k), \quad \frac{1}{n}\sum_{k=1}^{n} (g_f^k)^2(z_k) \geq \frac{1}{n}\sum_{k=1}^{n} (g_{z_{1:n},j_f}^k)^2(z_k) - 2\delta,$$

and therefore the probability of the set in (2.45) is bounded by

$$\mathcal{N}_1(\delta, \{\mathbf{g}_f : f \in \mathcal{F}\}, z_{1:n}) \max_{j \in \{1,\dots,N(z_{1:n})\}} \mathbb{P}\left(\{\delta + \frac{1}{n}\sum_{k=1}^{n} U_k g_{z_{1:n},j}^k(z_k) \geq c_1 - 2c_2\delta + c_2\frac{1}{n}\sum_{k=1}^{n} (g_{z_{1:n},j}^k)^2(z_k)\}\right).$$

Recast the values of c_0, c_1, c_2 in (2.43) and choose $\delta = \delta_0$ in the above as

$$\delta_0 = \frac{c_0 \alpha \rho}{2c_0(1-\varepsilon) + 1} \tag{2.46}$$

(where $\rho \in (0,1)$ is fixed in the hypotheses) to get the next partial result towards the proof of Theorem 2.3:

Lemma 2.13. *Under the hypothesis of Theorem 2.3 (with $B = 1$), with the notations from Section 2.2.1 and in (2.28), for any $n \in \mathbb{N}$, for any sequence $\mathbf{U} = (U_1, \dots, U_n)$ of independent Rademacher random variables independent of $Z_{1:n}$ and $Z'_{1:n}$, and for c_0, c_1, c_2, δ_0 given by (2.43), (2.46), let*

$$\{\mathbf{g}_{Z_{1:n},1}, \dots, \mathbf{g}_{z_{1:n},\mathcal{N}_1(\delta_0,\{\mathbf{g}_f : f \in \mathcal{F}\},Z_{1:n})}\}$$

be a (minimal) cover of $\{\mathbf{g}_f : f \in \mathcal{F}\}$ with respect to the empirical L_1 norm at $Z_{1:n}$. Then

$$\begin{aligned}
&\widetilde{\mathbb{P}}_n\left(\bigcup_{f \in \mathcal{F}} \{\widetilde{A}_n \mathbf{U}\mathbf{g}_f \geq c_1 + c_2 \widetilde{A}_n \mathbf{g}_f^2\}\right) \\
&\leq \mathcal{N}_1(\delta_0, \{\mathbf{g}_f : f \in \mathcal{F}\}, Z_{1:n}) \max_j \widetilde{\mathbb{P}}_n\left(\widetilde{A}_n \mathbf{U}\mathbf{g}_{Z_{1:n},j} \geq c_0(\alpha(1-\rho) + (1-\varepsilon)\widetilde{A}_n \mathbf{g}_{Z_{1:n},j}^2)\right).
\end{aligned} \tag{2.47}$$

▷ *Application of Bernstein's Inequality.* Consider the (random) numbers inside the maximum in (2.47):

$$\widetilde{\mathbb{P}}_n\left(\widetilde{A}_n\mathbf{U}\mathbf{g}_{Z_{1:n},j} \geq c_0(\alpha(1-\rho)+(1-\varepsilon)\widetilde{A}_n\mathbf{g}^2_{Z_{1:n},j})\right), \tag{2.48}$$

and denote by $\widetilde{\mathbb{V}}\mathrm{ar}_n\left[\,\cdot\,\right]$ the conditional variance given $Z_{1:n}$. Then since

$$\frac{1}{n}\sum_{k=1}^{n}\widetilde{\mathbb{V}}\mathrm{ar}_n\left[U_k g^k_{f_j}(Z_k)\right] = \widetilde{A}_n\mathbf{g}^2_{Z_{1:n},j},$$

the number in (2.48) has, for fixed $Z_{1:n}=z_{1:n}$, the form

$$\mathbb{P}\left(\frac{1}{n}\sum_{k=1}^{n}V_k \geq a_1 + a_2\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}V_k^2\right) \tag{2.49}$$

where $V_1,\dots,V_n$ are independent centered random variables with $|V_k|\leq 1$ $(k=1,\dots,n)$ and $a_1>0$, $a_2\geq 0$. The one-sided Bernstein's inequality (see the proof of [GKKW02, Lemma A.2]) gives the following bound for (2.49): if $\sigma_n^2 := \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}V_k^2$, then

$$\mathbb{P}\left(\frac{1}{n}\sum_{k=1}^{n}V_k \geq a_1+a_2\sigma_n^2\right) \leq \exp\left(-\frac{n(a_1+a_2\sigma_n^2)^2}{2\sigma_n^2+\frac{4}{3}(a_1+a_2\sigma_n^2)}\right) = \exp\left(-\frac{3na_2}{4}\frac{(\frac{a_1}{a_2}+\sigma_n^2)^2}{\frac{a_1}{a_2}+(\frac{3}{2a_2}+1)\sigma_n^2}\right)$$

$$\leq \exp\left(-\frac{3na_2}{4}\times 4\times\frac{a_1}{a_2}\times\frac{\frac{3}{2a_2}}{(\frac{3}{2a_2}+1)^2}\right) = \exp\left(-18\frac{na_1a_2}{(2a_2+3)^2}\right),$$

where in the third inequality we used the fact that

$$\frac{(a+u)^2}{a+bu} \geq 4a\frac{b-1}{b^2}$$

holds for $(a,b,u)\in(0,\infty)\times(0,\infty)\times(0,\infty)$.

Take now, as in (2.48), $a_1 = c_0\alpha(1-\rho)$ and $a_2 = c_0(1-\varepsilon)$ with c_0 as in (2.43). We obtain the following $\mathbb{P}-$a.s. upper bound

Lemma 2.14. *Under the same hypothesis and notation as in Lemma 2.13*

$$\widetilde{\mathbb{P}}_n\left(\widetilde{A}_n\mathbf{U}\mathbf{g}_{Z_{1:n},j} \geq c_0((1-\rho)\alpha+(1-\varepsilon)\widetilde{A}_n\mathbf{g}^2_{Z_{1:n},j})\right)$$
$$\leq \exp\left(-\frac{1}{2}\frac{(1-\varepsilon)}{(\frac{1}{3}(1-\frac{1}{\gamma'})\varepsilon(1-\varepsilon)+(1+\varepsilon))^2}\varepsilon^2(1-\frac{1}{\gamma'})^2(1-\rho)\alpha n\right)$$

for every $j\in\{1,\dots,\mathcal{N}_1(\delta_0,\{\mathbf{g}_f : f\in\mathcal{F}\},Z_{1:n})\}$.

▷ *Conclusion with "native" covering families (for $B=1$).* The analysis so far can be summarized as follows: putting together Lemmas 2.10 to 2.14 we arrive at the following result:

Lemma 2.15. *Under the hypothesis of Theorem 2.3 (with $B=1$), with the notations from Section 2.2.1, we have*

$$\mathbb{P}\left(\bigcup_{f\in\mathcal{F}}\{u(\widetilde{\mu_n}\mathbf{g}_f - \widetilde{A}_n\mathbf{g}_f) \geq \varepsilon(\alpha+\widetilde{\mu_n}\mathbf{g}_f)\}\right)$$

$$\leq \left(2\frac{\gamma}{\gamma-1}\right)^2 \mathbb{E}\left[\mathcal{N}_1(\frac{1}{2}\frac{(c-1)}{(c-\varepsilon)}\frac{1}{\gamma'}\alpha\varepsilon, \{\mathbf{g}_f^2 : f\in\mathcal{F}\}, Z_{1:n})\right]$$
$$\times \exp\left(-\frac{1}{2}\frac{1}{(1-\frac{1}{2}(1+\frac{1}{c})\varepsilon)^2}(1-\frac{1}{c})^2(1-\frac{1}{\gamma'})\varepsilon^2\alpha n\right)$$
$$+ 2\frac{\gamma}{\gamma-1}\mathbb{E}\left[\mathcal{N}_1(\frac{1}{2}\frac{1}{(1-\frac{1}{\gamma'})\varepsilon(1-\varepsilon)+(1+\varepsilon)}(1-\frac{1}{\gamma'})\alpha\varepsilon\rho, \{\mathbf{g}_f : f\in\mathcal{F}\}, Z_{1:n})\right]$$
$$\times \exp\left(-\frac{1}{2}\frac{(1-\varepsilon)}{(\frac{1}{3}(1-\frac{1}{\gamma'})\varepsilon(1-\varepsilon)+(1+\varepsilon))^2}(1-\frac{1}{\gamma'})^2(1-\rho)\varepsilon^2\alpha n\right), \tag{2.50}$$

provided that n satisfies (2.7).

The arguments that follow, which deal with the covering numbers appearing in (2.50) (an inequality valid for $B=1$) and with the conclusion for general $B>0$, can be explained appealing only to Lemma 2.15, thus we leave them as the "conclusive" step.

▷ *Homogenization of covering families and conclusion.* We have proved (2.50) under the hypothesis in Theorem 2.3 strengthened further to the assumption $B=1$. In order to get a bound of the right hand side of (2.50) that involves only covering numbers of the form $\mathcal{N}(\delta,\mathcal{F},X_{1:n})$, we note that if $h:\mathbb{R}^d\to\mathbb{R}$ is any Borel measurable function, then for any $f\in\mathcal{F}$ and any sequence $z=(z_k)_k=((x_k,y_k))_k$ of elements of $\mathbb{R}^d\times\mathbb{R}$, the hypothesis

$$\max\{|f(x_k)-y_k|, |h(x_k)-y_k|, |y_k-\Phi_k(x_k)|\}\leq 1$$

for all k gives:

$$\begin{aligned}|(g_h^k)^2(z_k)-(g_f^k)^2(z_k)| &= |g_h^k(z_k)-g_f^k(z_k)||g_h^k(z_k)+g_f^k(z_k)| \leq 2|g_h^k(z_k)-g_f^k(z_k)|\\ &= 2|(h(x_k)-y_k)^2-(\Phi_k(x_k)-y_k)^2-(f(x_k)-y_k)^2+(\Phi_k(x_k)-y_k)^2|\\ &= 2|h(x_k)+f(x_k)-2y_k||f(x_k)-h(x_k)| \leq 4|f(x_k)-h(x_k)|. \end{aligned}\tag{2.51}$$

Now, under the hypothesis

$$|f(x_k)-y_k|+|\Phi_k(x_k)-y_k|\leq 1$$

for every k and $f\in\mathcal{F}$ (which is true for $\mathbb{P}-$a.e. realization of (X_k,Y_k), according to (2.5) and the assumption $B=1$), every $\delta-L^1$cover $h_1,\dots,h_r$ of $\mathcal{F}$ with respect to the empirical norm

$$\frac{1}{n}\sum_{k=1}^n \delta_{\{x_k\}} \tag{2.52}$$

can be assumed to satisfy

$$|h_j(x_k)-y_k|+|y_k-\Phi_k(x_k)|\leq 1. \tag{2.53}$$

Indeed, according again to (2.5),

$$f(x_k)\in[y_k-1+|\Phi_k(x_k)-y_k|, y_k+1-|\Phi_k(x_k)-y_k|] =: [a_k,b_k]$$

for all $f\in\mathcal{F}$, so that given a $\delta-L^1$ cover $h_1,\dots,h_r$ of $\mathcal{F}$ with respect to (2.52) we can (if necessary) redefine

$$\hat{h}_r(x) = h_r(x)\mathbf{1}_{(a_k,b_k)}(h_r(x_k)) + a_k\mathbf{1}_{(-\infty,a_k]}(h_r(x_k)) + b_k\mathbf{1}_{[b_k,\infty)}(h_r(x_k))$$

to obtain a $\delta - L^1$ cover $\hat{h}_1, \ldots, \hat{h}_r$ of $\mathcal{F}$ (with respect to (2.52)) satisfying (2.53).

The inequalities in (2.51) imply therefore that for every $(z_k)_k = ((x_k, y_k))_k$ in $\mathbb{R}^d \times \mathbb{R}$ and every $\delta > 0$

$$\mathcal{N}_1(\delta, \{\mathbf{g}_f^2 : f \in \mathcal{F}\}, Z_{1:n}) \leq \mathcal{N}_1(\frac{\delta}{2}, \{\mathbf{g}_f : f \in \mathcal{F}\}, Z_{1:n}) \leq \mathcal{N}_1(\frac{\delta}{4}, \mathcal{F}, X_{1:n}), \tag{2.54}$$

$\mathbb{P}-$a.s.

The conclusion (2.6) for the case $B = 1$ follows from (2.50) and (2.54).

To finish the proof we consider the case in which $B > 0$ is arbitrary: we proceed as before by first dividing the event inequality in the left-hand side of (2.6) by B^2, applying the case $B = 1$ with α replaced by α/B^2, $\mathcal{F}$ by $\mathcal{F}/B$, and using (2.26). □

3 Applications to least–squares nonparametric regression

We develop in this section some applications of the previous results to the estimation of nonasymptotic errors for least–squares regression schemes, and we explore some of the consequences of these estimations with respect to the problem of consistency, both in the weak and strong senses, for the method of least–squares regression. See again sections 1.1 and 1.4 for general notation and conventions.

3.1 An $L^2_{\mathbb{P}}$-weak error estimate (Theorem 3.1)

A consequence of the estimates in the previous theorems is a control of the expected $L^2_{\mathbb{P}}$-deviation of Regression Averages. This is a generalization of [GKKW02, Theorem 11.5] with significant improvements.

Theorem 3.1. *Assume that $\mathcal{F}$ is a pointwise measurable family of functions $\mathbb{R}^d \to \mathbb{R}$ with associated Vapnik-Chervonenkis (VC) dimension $V_{\mathcal{F}} < +\infty$ and assume that, for some $B > 0$, $||Y_k||_{\mathbb{P},\infty} \leq B$ for all k. Then for $(c, \lambda) \in (1, \infty) \times (1, \infty)$ satisfying the hypothesis*

$$\lambda \leq \frac{3 + \sqrt{1 + 8c}}{4}, \qquad n \geq \exp(\frac{c^2 - 71}{4V_{\mathcal{F}}}), \tag{3.1}$$

the estimate

$$\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n}\int_{\mathbb{R}^d} |T_B\widehat{\Phi}_n(x) - \Phi_k(x)|^2 \mathbb{P}_{X_k}(\mathrm{d}x)\right] \leq \frac{B^2}{n}\theta_0\left(1 + \theta_1 + V_{\mathcal{F}}(\theta_2 + \log(\theta_2))\right)$$
$$+ \lambda \inf_{f \in \mathcal{F}} \frac{1}{n}\sum_{k=1}^{n}\int_{\mathbb{R}^d} |f(x) - \Phi_k(x)|^2 \mathbb{P}_{X_k}(\mathrm{d}x) \tag{3.2}$$

holds, where

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

The necessity for the truncation $T_B\widehat{\Phi}_n$ in (3.2) is discussed in details in Section 3.3.1. We point out that the posibility of choosing λ close to 1 depending on n will play an important role in Section 3.3 when dealing with consistency results.

Remark 3.2. This estimate improves the one in [GKKW02, Theorem 11.5]. For the sake of illustration, it is proved there that, in the i.i.d. case, and for $B \geq 1$, one has a bound of the type

$$\mathbb{E}\left[\int_{\mathbb{R}^d} |T_B\widehat{\Phi}_n(x) - \mathbb{E}\left[Y|X=x\right]|^2 \mathbb{P}_X(\mathrm{d}x)\right] \leq \frac{B^4}{n}\left(c_1 + V_{\mathcal{F}}(c_2 + c_3 \log n)\right) + 2 \inf_{f\in\mathcal{F}} \int_{\mathbb{R}^d} |f(x) - \mathbb{E}\left[Y|X=x\right]|^2 \mathbb{P}_X(\mathrm{d}x),$$

where $c_1 \geq 24332$, $c_2 \geq 73689$ and $c_3 \geq 10272$.

The use of $(\lambda, c) = (2, \sqrt{71})$ leads, for us, to a similar bound with B^4 replaced by B^2 (a gain due to the argument of homogeneization) and with c_1, c_2, and c_3 replaced respectively by constants c_1', c_2', and c_3' satisfying $c_1' \leq 7429$, $c_2' \leq 5967$ and $c_3' \leq 1852$. We can improve further the estimation of this deviation by plugging the greatest possible c according to (3.1), namely $c(n) := (4V_{\mathcal{F}} \log n + 71)^{1/2}$, into (3.2).

3.2 Proof of Theorem 3.1

We will use again the character "y" as in (2.4), and when necessary we implicitly interpret functions $f : \mathbb{R}^d \to \mathbb{R}$ as functions $\mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ evaluating in the first coordinate. The rest of the notation is also borrowed from the previous sections.

▷ *"Large deviation-bias" estimate.* Fix $\lambda > 1$, and consider the following identities

$$\begin{aligned}\mu_n|T_B\widehat{\mathbf{\Phi}}_n - \mathbf{\Phi}_{1:n}|^2 &= \widetilde{\mu}_n|T_B\widehat{\mathbf{\Phi}}_n - \mathbf{y}_{1:n}|^2 - \widetilde{\mu}_n|\mathbf{\Phi}_{1:n} - \mathbf{y}_{1:n}|^2 \\ &= \widetilde{\mu}_n \mathbf{g}_{T_B\widehat{\Phi}_n} - \lambda\widetilde{A}_n \mathbf{g}_{T_B\widehat{\Phi}_n} + \lambda\widetilde{A}_n \mathbf{g}_{T_B\widehat{\Phi}_n}, \end{aligned} \tag{3.3}$$

where $\widehat{\mathbf{\Phi}}_\mathbf{n} = (\widehat{\Phi}_n, \dots, \widehat{\Phi}_n)$ as in our convention. Now, by definition of $T_B\widehat{\Phi}_n$, the equation

$$\widetilde{A}_n \mathbf{g}_{T_B\widehat{\Phi}_n} = \widetilde{A}_n|T_B\widehat{\mathbf{\Phi}}_n - \mathbf{y}_{1:n}|^2 - \widetilde{A}_n|\mathbf{\Phi}_{1:n} - \mathbf{y}_{1:n}|^2 \leq \widetilde{A}_n|\widehat{\mathbf{\Phi}}_n - \mathbf{y}_{1:n}|^2 - \widetilde{A}_n|\mathbf{\Phi}_{1:n} - \mathbf{y}_{1:n}|^2 \leq \widetilde{A}_n \mathbf{g}_f, \tag{3.4}$$

holds for all $f \in \mathcal{F}$. Thus by the definition of $\mathbf{\Phi}_{1:n}$

$$\mathbb{E}\left[\widetilde{A}_n \mathbf{g}_{T_B\widehat{\Phi}_n}\right] \leq \inf_{f\in\mathcal{F}} \widetilde{\mu_n}\mathbf{g}_f = \inf_{f\in\mathcal{F}} \widetilde{\mu}_n(|\mathbf{f} - \mathbf{y}_{1:n}|^2 - |\mathbf{\Phi}_{1:n} - \mathbf{y}_{1:n}|^2) = \inf_{f\in\mathcal{F}} \mu_n|\mathbf{f} - \mathbf{\Phi}_{1:n}|^2. \tag{3.5}$$

Together, (3.3) and (3.5) imply, by integration, the "large deviation-bias" bound

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^n \int_{\mathbb{R}^d} |T_B\widehat{\Phi}_n(x) - \Phi_k(x)|^2 \mathbb{P}_{X_k}(\mathrm{d}x)\right] &:= \mathbb{E}\left[\mu_n|T_B\widehat{\mathbf{\Phi}}_n - \mathbf{\Phi}_{1:n}|^2\right] \\ &\leq \mathbb{E}\left[\widetilde{\mu}_n \mathbf{g}_{T_B\widehat{\Phi}_n} - \lambda\widetilde{A}_n \mathbf{g}_{T_B\widehat{\Phi}_n}\right] + \lambda \inf_{f\in\mathcal{F}} \mu_n|\mathbf{f} - \mathbf{\Phi}_{1:n}|^2.\end{aligned}$$

▷ *Bounding the large deviation term.* Now we bound the cumulative distribution function of $\widetilde{\mu}_n \mathbf{g}_{T_B\widehat{\Phi}_n} - \lambda\tilde{A}_n \mathbf{g}_{T_B\widehat{\Phi}_n}$. To do so we will assume first that $B = 1/4$. The goal of this choice is to use for simplicity (2.5) and its conclusion (2.6) with $B = 1$.

We start by noticing that, if $T_{1/4}\mathcal{F} := \{T_{1/4}f : f \in \mathcal{F}\}$ then

$$\begin{aligned}\left\{\widetilde{\mu}_n \mathbf{g}_{T_{1/4}\widehat{\Phi}_n} - \lambda \widetilde{A}_n \mathbf{g}_{T_{1/4}\widehat{\Phi}_n} \geq t\right\} &\subset \bigcup_{f \in T_{1/4}\mathcal{F}} \left\{\widetilde{\mu_n}\mathbf{g}_f - \lambda \widetilde{A}_n \mathbf{g}_f \geq t\right\} \\ &= \bigcup_{f \in T_{1/4}\mathcal{F}} \left\{\widetilde{\mu_n}\mathbf{g}_f - \widetilde{A}_n \mathbf{g}_f \geq \frac{(\lambda-1)}{\lambda}\left(\frac{t}{\lambda-1} + \widetilde{\mu_n}\mathbf{g}_f\right)\right\}.\end{aligned} \tag{3.6}$$

We apply at this step the inequality (2.6) with the parameters

$$(B, \alpha, \varepsilon, \rho, \gamma, \gamma') = (1, \frac{t}{\lambda-1}, 1 - \frac{1}{\lambda}, \frac{1}{c}, 1 + \frac{1}{c}, c) \tag{3.7}$$

to get an estimate of the form

$$\begin{aligned}\mathbb{P}\left(\widetilde{\mu}_n \mathbf{g}_{T_{1/4}\widehat{\Phi}_n} - \lambda \widetilde{A}_n \mathbf{g}_{T_{1/4}\widehat{\Phi}_n} \geq t\right) &\leq G_0(c)\mathbb{E}\left[\mathcal{N}_1(G_1(c,\lambda)t, T_{1/4}\mathcal{F}, X_{1:n})\right]\exp(-b_{c,\lambda}nt) \\ &=: a_{c,\lambda}(t)\exp(-b_{c,\lambda}nt)\end{aligned} \tag{3.8}$$

where

$$\begin{aligned}G_0(c) &:= 2(c+1)(2c+3), \qquad G_1(c,\lambda) := \frac{1}{8}\frac{1}{\lambda(c-1)+1}(1-\frac{1}{c}), \\ b_{c,\lambda} &:= \frac{1}{2}\frac{1}{(\frac{1}{3}(1-\frac{1}{c}) + (2\lambda-1)\frac{\lambda}{\lambda-1})^2}(1-\frac{1}{c})^3\frac{\lambda}{\lambda-1}.\end{aligned} \tag{3.9}$$

The verification of (3.8) with (3.9) is given later (see page 28). By (2.7) and (3.7) this estimate is valid for $t > 0$ provided that

$$n \geq \frac{c(c+1)\lambda^2}{4t(\lambda-1+t)} \Leftrightarrow t \geq \frac{-(\lambda-1) + \sqrt{(\lambda-1)^2 + c(c+1)\lambda^2/n}}{2} =: t_n(c,\lambda). \tag{3.10}$$

Since covering numbers are decreasing with respect to the radius of the covering, this gives rise to the inequality

$$\mathbb{P}\left(\widetilde{\mu}_n \mathbf{g}_{T_{1/4}\widehat{\Phi}_n} - \lambda \widetilde{A}_n \mathbf{g}_{T_{1/4}\widehat{\Phi}_n} \geq t\right) \leq a(t_n)\exp(-bnt), \tag{3.11}$$

for $t \geq t_n$ and where, for simplicity, we drop the parameters c, λ from the notation for a, b and t_n.

Note now that the function $a(t_n)$ here depends on the distribution of $X_{1:n}$. In order to bound it in a distribution-free way, we start by noticing that

$$T_{1/4}\mathcal{F} + 1/4 := \{f + 1/4, f \in T_{1/4}\mathcal{F}\}$$

is a family of nonnegative functions bounded by $1/2$ which admits the same covering numbers as $T_{1/4}\mathcal{F}$ with respect to the empirical measure associated to $x_{1:n} := \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$. Additionally we observe that $G_1 t_n < 1/8$. Indeed, since $c > 1$ we obtain, via the inequality $\sqrt{a^2+b^2} \leq |a| + |b|$, that

$$\begin{aligned}8G_1 t_n &= \frac{1}{2c}\frac{c-1}{(c-1+\frac{1}{\lambda})}\left(-(1-\frac{1}{\lambda}) + \sqrt{(1-\frac{1}{\lambda})^2 + \frac{c(c+1)}{n}}\right) \leq \frac{1}{2}\sqrt{\frac{(1+\frac{1}{c})}{n}} \\ &\leq \left(\frac{1}{2n}\right)^{1/2} < 1.\end{aligned} \tag{3.12}$$

We apply at this step the estimates in [GKKW02, Lemma 9.2 and Theorem 9.4] and the fact that $V_{T_{1/4}\mathcal{F}+1/4} = V_{T_{1/4}\mathcal{F}} \leq V_{\mathcal{F}}$ to conclude that

$$\mathcal{N}_1(G_1 t_n, T_{1/4}\mathcal{F}, x_{1:n}) = \mathcal{N}_1(G_1 t_n, T_{1/4}\mathcal{F} + \frac{1}{4}, x_{1:n}) \leq 3\left(\frac{e}{G_1 t_n}\log(\frac{3e}{2G_1 t_n})\right)^{V_{\mathcal{F}}}$$

(note that these estimates hold also in the trivial case in which $V_{T_{1/4}\mathcal{F}} = 1$, excluded by the statement of [GKKW02, Theorem 9.4]). These inequalities bring us to the following (distribution-free) estimate of the (distribution-dependent) coefficient $a(t_n)$ of the exponential function at the right-hand side of (3.11):

$$\begin{aligned} a(t_n) \equiv a_{c,\lambda}(t_n) &:= G_0(c)\mathbb{E}\left[\mathcal{N}_1(G_1(c,\lambda)t_n, T_{1/4}\mathcal{F}, X_{1:n}]\right) \\ &\leq 3G_0(c)\left(\frac{e}{G_1(c,\lambda)t_n}\log(\frac{3e}{2G_1(c,\lambda)t_n})\right)^{V_{\mathcal{F}}} := G(t_n, V_{\mathcal{F}}). \end{aligned} \tag{3.13}$$

Appying Lemma 2.1 to (3.11) combined with (3.13), we get that for t_n as in (3.10) and for the ($(\lambda, c)-$dependent) objects in (3.9), the hypothesis

$$3G_0\left(\frac{e}{G_1 t_n}\log(\frac{3}{2}\frac{e}{G_1 t_n})\right)^{V_{\mathcal{F}}} \geq \exp(bnt_n) \tag{3.14}$$

implies that

$$\mathbb{E}\left[\widetilde{\mu}_n \mathbf{g}_{T_{1/4}\widehat{\Phi}_n} - \lambda \widetilde{A}_n \mathbf{g}_{T_{1/4}\widehat{\Phi}_n}\right] \leq \frac{1}{bn}\left(1 + \log\left(3G_0(\frac{e}{G_1 t_n}\log(\frac{3}{2}\frac{e}{G_1 t_n}))^{V_{\mathcal{F}}}\right)\right). \tag{3.15}$$

▷ *Verification of* (3.14) *under* (3.1). Since the (bijective) function $[0,\infty) \to [0, a/2)$ defined by

$$x \mapsto \sqrt{x^2 + ax} - x$$

is increasing, it follows that

$$nt_n = \frac{\lambda - 1}{2}\left(\sqrt{n^2 + c(c+1)\frac{\lambda^2}{(\lambda-1)^2}n} - n\right) \leq \frac{c(c+1)}{4}\frac{\lambda^2}{(\lambda-1)}, \tag{3.16}$$

and therefore, using the expression for $b = b_{c,\lambda}$ in (3.9),

$$bnt_n \leq \frac{1}{8}\frac{1}{(\frac{1}{3}(1-\frac{1}{c})(1-\frac{1}{\lambda}) + (2\lambda - 1))^2}(1 - \frac{1}{c})^2(c^2-1)\lambda \leq \frac{c^2-1}{8\lambda} < \frac{c^2-1}{8} \tag{3.17}$$

(note that these bounds do not depend on n).

In order to guarantee (3.14) we note the following: first, the string of inequalities (3.12) gives the bound

$$G_1 t_n \leq (2^7 n)^{-1/2}. \tag{3.18}$$

Therefore a sufficient condition for (3.14) is, by (3.17) and (3.18), the following

$$\log(3G_0) + V_{\mathcal{F}}(1 + \frac{1}{2}(\log(2^7) + \log(n)) + \log(1 + \log(3) + \frac{1}{2}(\log(2^5) + \log(n))) \geq \frac{c^2-1}{8},$$

which, using $8[\log(3G_0) + V_{\mathcal{F}}(1 + \frac{1}{2}\log(2^7) + \log(1 + \log(3) + \frac{1}{2}\log(2^5)))] \geq 8[\log(60) + (1 + \frac{1}{2}\log(2^7) + \log(1 + \log(3) + \frac{1}{2}\log(2^5)))] \geq 70$, clearly holds if

$$\log n \geq \frac{c^2 - 71}{4V_{\mathcal{F}}}.$$

▷ *An upper bound for the right hand side of* (3.15). Note now that from the first equality in (3.12) and the fact that the function $[0, \infty) \to [0, \sqrt{a}]$ defined by

$$x \mapsto \sqrt{x^2 + a} - x$$

is decreasing it follows that

$$8G_1 t_n \geq \frac{1}{2c}(1 - \frac{1}{c})\left(-1 + \sqrt{1 + \frac{c(c+1)}{n}}\right) = \frac{1}{2c}\frac{(c^2-1)}{n}\frac{1}{1 + \sqrt{1 + \frac{c(c+1)}{n}}}. \tag{3.19}$$

This implies by an easy argument the bound

$$\log(3G_0\left(\frac{e}{G_1 t_n}\log(\frac{3}{2}\frac{e}{G_1 t_n})\right)^{V_{\mathcal{F}}}) \leq \log(3G_0) + V_{\mathcal{F}}(\theta_2 + \log(\theta_2))$$

with

$$\theta_2 = \theta_2(n, c) := 1 + \log 24 + \log(1 + \sqrt{1 + \frac{c(c+1)}{n}}) - \log(c - \frac{1}{c}) + \log n.$$

All the elements of the proof are gathered for the case $B = 1/4$.

▷ *The case of arbitrary* $B > 0$. This follows from the following homogenization properties: for any $((x_k, y_k))_{k=1}^n$, and defining $\mathcal{F}/4B$ as in (2.25)

- $\int |f(x) - \Phi_k(x)|^2 \mathbb{P}_{X_k}(\mathrm{d}x) = (4B)^2 \int |\frac{f(x)}{4B} - \frac{\Phi_k(x)}{4B}|^2 \mathbb{P}_{X_k}(\mathrm{d}x)$,
- $f^* \in \arg\min_{f \in \mathcal{F}} \frac{1}{n}\sum_{k=1}^n |f(x_k) - y_k|^2 \Leftrightarrow \frac{f^*}{4B} \in \arg\min_{f \in \frac{\mathcal{F}}{4B}} \frac{1}{n}\sum_{k=1}^n |f(x_k) - \frac{y_k}{4B}|^2$,
- $V_{\frac{\mathcal{F}}{4B}} = V_{\mathcal{F}}$,

from where one deduces the estimate for arbitrary $B > 0$ via the identity

$$\widetilde{\mu}_n |T_B \widehat{\mathbf{\Phi}}_n - \mathbf{\Phi}_{1:n}|^2 - \lambda \inf_{f \in \mathcal{F}} \widetilde{\mu}_n |\mathbf{f} - \mathbf{\Phi}_{1:n}|^2 = (4B)^2 \left(\widetilde{\mu}_n |T_{1/4} \frac{\widehat{\mathbf{\Phi}}_n}{4B} - \frac{\mathbf{\Phi}_{1:n}}{4B}|^2 - \lambda \inf_{f \in \frac{\mathcal{F}}{4B}} \widetilde{\mu}_n |\mathbf{f} - \frac{\mathbf{\Phi}_{1:n}}{4B}|^2\right)$$

and the case $B = 1/4$ already treated.

▷ *Proof of the inequality* (3.8) *with the constants* (3.9). Applying the inequality (2.6) to bound the probability of the right-hand-side of (3.6) with the parameters (3.7) gives the bound

$$\begin{aligned}\mathbb{P}\left(\widetilde{\mu}_n \mathbf{g}_{T_{1/4}\widehat{\Phi}_n} - \lambda \widetilde{A}_n \mathbf{g}_{T_{1/4}\widehat{\Phi}_n} \geq t\right) \leq (a(c))^2 \mathbb{E}\left[\mathcal{N}_1(R_1(c,\lambda)t, T_{1/4}\mathcal{F}, X_{1:n})\right] \exp(-b_1(c,\lambda)nt) \\ + a(c)\mathbb{E}\left[\mathcal{N}_1(R_2(c,\lambda)t, T_{1/4}\mathcal{F}, X_{1:n})\right] \exp(-b_2(c,\lambda)nt)\end{aligned}$$

where

$$a(c) := 2(c+1),$$

$$R_1(c,\lambda) := \frac{1}{8}\frac{1}{\lambda(c-1)+1}(1-\frac{1}{c}), \qquad b_1(c,\lambda) := \frac{1}{2}\frac{1}{(\frac{\lambda}{\lambda-1}-\frac{1}{2}(1+\frac{1}{c}))^2}(1-\frac{1}{c})^3\frac{1}{\lambda-1},$$
$$R_2(c,\lambda) := \frac{1}{4}\frac{1}{c(2\lambda-\frac{1}{\lambda})-(1-\frac{1}{\lambda})}(1-\frac{1}{c}), \quad b_2(c,\lambda) := \frac{1}{2}\frac{1}{(\frac{1}{3}(1-\frac{1}{c})+(2\lambda-1)\frac{\lambda}{\lambda-1})^2}(1-\frac{1}{c})^3\frac{\lambda}{\lambda-1}.$$

Now, an elementary computation shows that, for $(\lambda, c) \in (1,\infty)\times[1,\infty)$, $b_1(c,\lambda) \geq b_2(c,\lambda)$, and that the condition $R_1(c,\lambda) \leq R_2(c,\lambda)$ is equivalent to $\lambda \leq (3+\sqrt{1+8c})/4$, which is part of condition (3.1). Therefore we have (always for $t \geq t_n$) that

$$\mathbb{P}\left(\widetilde{\mu}_n \mathbf{g}_{T_{1/4}\widehat{\Phi}_n} - \lambda \widetilde{A}_n \mathbf{g}_{T_{1/4}\widehat{\Phi}_n} \geq t\right) \leq a(c)(a(c)+1)\mathbb{E}\left[\mathcal{N}_1(R_1(c,\lambda)t, T_{1/4}\mathcal{F}, X_{1:n})\right]\exp(-b_2(c,\lambda)nt).$$

Putting together these pieces, we verify (3.8) with (3.9) as described above.

Remark 3.3 (Scope of our methods for the case of infinite complexity.)**.** In the proof of Theorem 3.1, we use a bound of the type (see Equation (3.13))

$$\log \mathcal{N}_1(r,\mathcal{F},x_{1:n}) \leq C_{V_\mathcal{F}} \log r^{-1}, \tag{3.20}$$

provided that r is small. Notice that (3.20) implies a bound of the type

$$\log \mathcal{N}_1(r,\mathcal{F},x_{1:n}) \leq C_\mathcal{F}\, r^{-\alpha}, \tag{3.21}$$

for every $\alpha > 0$ and $n \in \mathbb{N}$.

The assumption (3.21) (for $\alpha \in (0,2)$), among others, is used in [HW17], which addresses the problem of optimal rates for i.i.d. least–squares regression schemes with heavy–tailed distributions. Notice that if we are under (3.21) but not under (3.20), then necessarily $V_\mathcal{F} = \infty$ (consider the point in the proof of Theorem 3.1 in which (3.13) is invoked).

This leads to the natural question of whether we can give a rate for the case of infinite complexity (3.21) with our methods. The answer is affirmative at least when $\alpha \in (0,1)$: using (3.21) instead of (3.13) in the previous proof one can verify that, for $\alpha \in (0,1)$, we arrive at a function of the form $O(n^{\alpha-1})$ in the place of the first term at the right–hand side of (3.2) (the "variance" term), with "$O(\cdot)$" constants depending on $(\mathcal{F}, B, c, \lambda)$ that admit an explicit expression. The details are left to the reader.[4]

3.3 Consistency

The estimate (3.2) and its proof have relevant consequences for the problem of consistency of least–squares regression schemes, some of which we will describe in this section. In what follows, we often use for the sake of simplicity the short notations given in (1.7).

To begin with the discussion notice the following: according to Theorem 3.1, if both $\mathcal{F}$ ($V_\mathcal{F} < \infty$) and the response functions Y_k are uniformly bounded by B, then the least squares regression estimate is *weakly consistent* in the sense that

$$\lim_n \left(\mathbb{E}\left[\mu_n|\widehat{\mathbf{\Phi}}_n - \mathbf{\Phi}_{1:n}|^2\right] - \inf_{f\in\mathcal{F}} \mu_n|\mathbf{f}-\mathbf{\Phi}_{1:n}|^2\right) = 0, \tag{3.22}$$

[4]Note additionally that the rate obtained by intersecting the hypotheses in [HW17] with our hypotheses, and in particular assuming that (3.21) holds for every $\alpha > 0$, corresponds up to a logarithmic factor to the one obtained by letting $\alpha \to 0$ in the conclusion of [HW17, Theorem 4]. Since the rates in [HW17] are sharp in a general sense (under the hypotheses in that paper), this may be evidence that our variance bound in Theorem 3.1 has an (almost) sharp asymptotic rate for the setting considered here.

which follows easily from the fact that, in this case, $T_B\widehat{\Phi}_n = \widehat{\Phi}_n \in \mathcal{F}$ ($\mathcal{F}$ is a space of functions bounded by B) by substracting $\inf_{f\in\mathcal{F}} \mu_n|\mathbf{f}-\mathbf{\Phi}_{1:n}|^2 (\leq 4B^2)$ at both sides of (3.2) and letting $n\to\infty$ because $\lambda > 1$ is arbitrary. We will see later (see Corollary 3.13) that, in this context, (3.22) holds actually in the *strong* sense, i.e. the equality

$$\lim_n \left(\mu_n|\widehat{\mathbf{\Phi}}_n - \mathbf{\Phi}_{1:n}|^2 - \inf_{f\in\mathcal{F}} \mu_n|\mathbf{f}-\mathbf{\Phi}_{1:n}|^2 \right) = 0$$

holds with probability one.

3.3.1 The Role of Truncation for Consistency and Stability

Note that (3.22) follows from an interpretation in the present case of an inequality of the type

$$\mathbb{E}\left[\mu_n|T_B\widehat{\mathbf{\Phi}}_n - \mathbf{\Phi}_{1:n}|^2\right] \leq \delta(\lambda,n) + \lambda \inf_{f\in\mathcal{F}} \mu_n|\mathbf{f}-\mathbf{\Phi}_{1:n}|^2 \tag{3.23}$$

where $\delta : (1,\infty)\times\mathbb{N}\to(0,\infty)$ denotes a generic function with the property that $\delta(\lambda,n)\to_n 0$ for every fixed λ (sufficiently close to one). To discuss the role of the truncation operator in the inequality (3.23) in the context of Theorem 3.1 notice first that, under the hypotheses of this theorem, and for any $f\in\mathcal{F}$, the inequality

$$|T_B f - \Phi_k|^2 \leq |f-\Phi_k|^2$$

holds $\mathbb{P}_{X_k}-$a.s. for every k, and this implies that (3.23) holds provided that either of the following type of inequalities

$$\mathbb{E}\left[\mu_n|T_B\widehat{\mathbf{\Phi}}_n - \mathbf{\Phi}_{1:n}|^2\right] \leq \delta(\lambda,n) + \lambda \inf_{f\in\mathcal{F}} \mu_n|T_B\mathbf{f}-\mathbf{\Phi}_{1:n}|^2, \tag{3.24}$$

$$\mathbb{E}\left[\mu_n|\widehat{\mathbf{\Phi}}_n - \mathbf{\Phi}_{1:n}|^2\right] \leq \delta(\lambda,n) + \lambda \inf_{f\in\mathcal{F}} \mu_n|\mathbf{f}-\mathbf{\Phi}_{1:n}|^2 \tag{3.25}$$

hold. Notice also that, by the same argument leading to the proof of (3.22), the inequality (3.24) [resp. (3.25)] can be interpreted as an inequality stating the weak consistency of $T_B\widehat{\Phi}_n$[resp. $\widehat{\Phi}_n$] as an estimator of $T_B\operatorname{arginf}_{f\in\mathcal{F}}\mu_n|T_B\mathbf{f}-\mathbf{\Phi}_{1:n}|^2$ [resp. $\operatorname{arginf}_{f\in\mathcal{F}}\mu_n|\mathbf{f}-\mathbf{\Phi}_{1:n}|^2$].

But neither of the inequalities (3.24) or (3.25) can be derived solely from the hypotheses of Theorem 3.1, as we now justify.

▷ *Counterexample to* (3.24)*:* let $(X_k,Y_k) = (X,Y)$ be a unit mass at $(0,1)$ for every k ($((X_k,Y_k))_k$ is i.i.d.), so that $\mu_n = \mu := \mathbb{P}_X = \delta_0$ (a Dirac mass at zero) for every n and $\Phi_k(0) = 1$ for every k, and let $\mathcal{F} := \{f_1, f_2\}$ with $f_1 \equiv 1/2$ and $f_2 \equiv 2$. In this case, it is clear that $\widehat{\Phi}_n = f_1$ and that

$$\inf_{f\in\mathcal{F}} \mu_n|T_1\mathbf{f}-\mathbf{\Phi}_{1:n}|^2 = \inf_{f\in\mathcal{F}} |T_1 f(0)-1|^2 = |T_1 f_2(0)-1|^2 = 0,$$

and therefore, for this example

$$\mu_n|T_1\widehat{\mathbf{\Phi}}_n - \mathbf{\Phi}_{1:n}|^2 - \inf_{f\in\mathcal{F}} \mu_n|T_1\mathbf{f}-\mathbf{\Phi}_{1:n}|^2 = \mu_n|T_1\widehat{\mathbf{\Phi}}_n - \mathbf{\Phi}_{1:n}|^2 = 1/2 \tag{3.26}$$

for every n, which makes impossible a bound of the type (3.24).

▷ *Counterexample to* (3.25) Let X and Y be independent with X uniformly distributed on $[0,1]$ and Y Rademacher distributed ($\mathbb{P}(Y=1)=\mathbb{P}(Y=-1)=1/2$, notice that $\mathbb{E}[Y|X]=0$), let $((X_k,Y_k))_k$ be i.i.d. $(X_k,Y_k)\sim(X,Y)$, and let $\mathcal{F}$ be the (finite-dimensional) vector space of piecewise linear functions in a fixed finite partition by intervals of $[0,1]$. Then again $\mu_n=\mu:=\mathbb{P}_X$ and, as stated in [GKKW02, Problem 10.3],

$$\mathbb{E}\left[\mu_n|\widehat{\mathbf{\Phi}}_n-\mathbf{\Phi}_{1:n}|^2\right]=\infty \tag{3.27}$$

for every n, whereas

$$\inf_{f\in\mathcal{F}}\mu_n|\mathbf{f}-\mathbf{\Phi}_{1:n}|^2=0, \tag{3.28}$$

because, for all k, $\Phi_k\equiv 0\in\mathcal{F}$. Notice that (3.27) and (3.28) are an obstruction to (3.25).

The message from the previous examples is summarized in the following conclusions: first, *according to* (3.26), *the truncated least squares regressor* $T_B\widehat{\Phi}_n$ *is* not *(in general) a consistent estimator of* $\arg\min_{f\in\mathcal{F}}\mu_n|T_B\mathbf{f}-\mathbf{\Phi}_{1:n}|^2$; second, *according to* (3.27), *the non-truncated least–squares regression estimator* $\widehat{\Phi}_n$ *is* not *(in general) a (weakly-)consistent,* nor *an stable estimator of* $\arg\min_{f\in\mathcal{F}}\mu_n|\mathbf{f}-\mathbf{\Phi}_{1:n}|^2$; and third, *the bound* (2.6) *in Theorem 2.3 is optimal in terms of the use of the truncation operator.*

In the remaining of this section, we will derive some consistency results based on a more careful interplay between the truncation bounds, the sample sizes, and the function–space complexities (VC-dimensions) of the respective schemes.

3.3.2 The setting and the working hypotheses

The upcoming string of results will be given under the following setting, which we specify in advance to make more efficient the statements and to facilitate comparisons.

S 1 (Truncated LSR for row-Independent and response-bounded triangular-arrays)**.** *For every* $m\in\mathbb{N}$, *let* $D_m:=\{(X_{m,k},Y_{m,k})\}_{k=1}^{n_m}$ *($n_m\geq 3$) be an independent random sequence in* $\mathbb{R}^{d_m}\times[-B_m,B_m]$ *with associated regression functions*

$$\Phi_{m,k}(x)=\mathbb{E}[Y_{m,k}|X_{m,k}=x],\qquad \mathbb{P}_{X_{m,k}}-a.s.$$

Let $\mathcal{F}_m$ *be a family of functions* $\mathbb{R}^{d_m}\to\mathbb{R}$ *with* $V_{\mathcal{F}_m}<\infty$, *and let* $\widehat{\Phi}_{n_m}^{(m)}$ *be given by* (1.2) *with* $\mathcal{F}$ *replaced by* $\mathcal{F}_m$ *and* $\{(X_k,Y_k)\}_{k=1}^n$ *replaced by* D_m.

We also introduce the following hypotheses:

H 1. *Under (S1),*

$$\lim_m\left(\frac{V_{\mathcal{F}_m}B_m^2\log n_m}{n_m\delta_m}+\delta_m\inf_{f\in\mathcal{F}_m}\frac{1}{n_m}\sum_{k=1}^{n_m}\int_{\mathbb{R}^{d_m}}|f(x)-\Phi_{m,k}(x)|^2\mathbb{P}_{X_{m,k}}(\mathrm{d}x)\right)=0 \tag{3.29}$$

for some positive sequence $(\delta_m)_m$ *with* $\delta_m\to_m 0$.

and its stronger version

H 2. *Under (S1),*

$$\frac{V_{\mathcal{F}_m}B_m^2\log n_m}{n_m}\to_m 0,\qquad \limsup_m\inf_{f\in\mathcal{F}_m}\frac{1}{n_m}\sum_{k=1}^{n_m}\int_{\mathbb{R}^{d_m}}|f(x)-\Phi_{m,k}(x)|^2\mathbb{P}_{X_{m,k}}(\mathrm{d}x)<\infty.$$

Notice that the satisfaction of H1 implies that $n_m \to +\infty$ as $m \to +\infty$, except on sub-sequences where $B_m \to 0$ and in this last case, the results become somehow trivial. Therefore, in most usual cases, we implicitly have $n_m \to +\infty$ as $m \to +\infty$.

Remark 3.4. Note that, under the hypothesis

$$\limsup_m \inf_{f \in \mathcal{F}_m} \frac{1}{n_m} \sum_{k=1}^{n_m} \int_{\mathbb{R}^{d_m}} |f(x) - \Phi_{m,k}(x)|^2 \mathbb{P}_{X_{m,k}}(\mathrm{d}x) < \infty, \tag{3.30}$$

H1 and H2 are equivalent: they hold if and only if

$$\frac{V_{\mathcal{F}_m} B_m^2 \log n_m}{n_m} \to_m 0,$$

which can be seen by considering

$$\delta_m := \left(\frac{V_{\mathcal{F}_m} B_m^2 \log n_m}{n_m} \right)^\rho \tag{3.31}$$

for any $\rho \in (0,1)$. In particular H2 implies H1.

Note also that (3.30) can be given for granted for basically any conceivable application to statistical learning of the results that follow. We keep the separate hypothesis H1 and H2 because, as we shall see, H2 permits some specifications that are not explicit under H1.

Remark 3.5 (Extensions to infinite complexity. See also Remark 3.3)**.** When (3.21) holds for all m with $\mathcal{F}$ replaced by $\mathcal{F}_m$ and α replaced by $\alpha_m \in (0,1)$, and assuming without loss of generality that the corresponding $C_{\mathcal{F}_m}$ satisfies $C_{\mathcal{F}_m} \geq 1$, results similar to those that follow under H1 can be obtained under the condition obtained replacing

$$\frac{V_{\mathcal{F}_m} B_m^2 \log n_m}{n_m \delta_m} \tag{3.32}$$

by

$$\frac{C_{\mathcal{F}_m} B_m^2}{n_m^{1-\alpha_m} \delta_m} \tag{3.33}$$

in (3.29). The necessary modifications in the proofs and the statements will be easy to implement.

3.3.3 Weak Consistency (Theorem 3.6)

Our main result on weak consistency gives a condition on the relationship between the parameters $n_m, V_{\mathcal{F}_m}$, and B_m in order to guarantee that, on average, the deviation obtained from the truncated least squares regression function is, in the limit, smaller (or equal) than the deviation obtained by the best approximation within the respective spaces. Its proof is an immediate consequence of the bound in Theorem 3.1.

Theorem 3.6. *Assume H1, then*

$$\limsup_m \left(\mathbb{E}\left[\frac{1}{n_m} \sum_{k=1}^{n_m} \int_{\mathbb{R}^{d_m}} |T_{B_m} \widehat{\Phi}_{n_m}^{(m)}(x) - \Phi_{m,k}(x)|^2 \mathbb{P}_{X_{m,k}}(\mathrm{d}x) \right] - \inf_{f \in \mathcal{F}_m} \frac{1}{n_m} \sum_{k=1}^{n_m} \int_{\mathbb{R}^{d_m}} |f(x) - \Phi_{m,k}(x)|^2 \mathbb{P}_{X_{m,k}}(\mathrm{d}x) \right) \leq 0. \tag{3.34}$$

Proof. Choose $\lambda_m = 1 + \delta_m$ and let $c \geq 1$ be such that the hypotheses (3.1) of Theorem 3.1 hold for every m with the input $(n_m, B_m, \mathcal{F}_m, c, \lambda_m)$[5]. Then, according to (3.2),

$$\mathbb{E}\left[\mu_{n_m}^{(m)} |T_{B_m} \widehat{\mathbf{\Phi}}_{n_m}^{(m)} - \mathbf{\Phi}_{m,1:n_m}|^2\right] - \inf_{f \in \mathcal{F}_m} \mu_{n_m}^{(m)} |\mathbf{f} - \mathbf{\Phi}_{m,1:n_m}|^2$$
$$\leq \frac{B_m^2}{n_m} \theta_0^{(m)} \left(1 + \theta_1^{(m)} + V_{\mathcal{F}_m}(\theta_2^{(m)} + \log(\theta_2^{(m)}))\right) + \delta_m \inf_{f \in \mathcal{F}_m} \mu_{n_m}^{(m)} |\mathbf{f} - \mathbf{\Phi}_{m,1:n_m}|^2, \tag{3.35}$$

where we have set $\mu_{n_m}^{(m)} \mathbf{h} := \frac{1}{n_m} \sum_{k=1}^{n_m} \int_{\mathbb{R}^d} h_k(x) \mathbb{P}_{X_{m,k}}(\mathrm{d}x)$ as a natural extension of (1.7). The convergence to zero of the right hand side of (3.35) as $m \to \infty$ is then straightforward from (3.29), the details are left to the reader. □

Remark 3.7. Notice that the inequality (3.35) is nonasymptotic and can be interpreted as a rate of convergence of the error between the mean L^2 distance of the truncated least–squares regressor over $\mathcal{F}_m$ to the "true" regressor and the minimal of such L^2 distances (rate of "weak" convergence).

The reader is particularly invited to consider the meaning of (3.35) and its interactions with H1 in the i.i.d. case (i.e. $(X_k, Y_k) \sim (X, Y)$ for all k).[6]

Remark 3.8. Let (X, Y) be a point mass at $(0, 1)$, let $(X_k, Y_k) = (X, Y)$ for every k ($(X_k, Y_k)_{k \in \mathbb{N}}$ are i.i.d.), and let $\mathcal{F} := \{f\}$ where $f(0) = 2$. Specializing to this case with $B_m = 1$ for all m we see that the left hand side of (3.34) can be strictly negative: it is not possible to prove the equality reverse to (3.34) under H1.

An elementary consequence of Theorem 3.6 is the following generalization of (3.22) in which the sets of function hypotheses, the sample sizes, and the response bounds can increase following an explicit control, which allows to consider scenarios of increasing complexity (including regression schemes over asymptotically infinite–dimensional spaces).

Corollary 3.9. *Assume H1. If for every m, $\mathcal{F}_m$ is a space of functions uniformly bounded by B_m, then*

$$\lim_m \left(\mathbb{E}\left[\frac{1}{n_m} \sum_{k=1}^{n_m} \int_{\mathbb{R}^{dm}} |\widehat{\Phi}_{n_m}^{(m)}(x) - \Phi_{m,k}(x)|^2 \mathbb{P}_{X_{m,k}}(\mathrm{d}x)\right] - \right.$$
$$\left. \inf_{f \in \mathcal{F}_m} \frac{1}{n_m} \sum_{k=1}^{n_m} \int_{\mathbb{R}^{dm}} |f(x) - \Phi_{m,k}(x)|^2 \mathbb{P}_{X_{m,k}}(\mathrm{d}x) \right) = 0. \tag{3.36}$$

In particular, (3.36) holds if, under the setting S1,

$$\frac{V_{\mathcal{F}_m} B_m^4 \log n_m}{n_m} \to_m 0. \tag{3.37}$$

Proof. Note first that the $\liminf_m$ of the numbers inside the $\lim_m$ in (3.36) is nonnegative. The $\limsup_m$ is nonpositive according to Theorem 3.6 assuming that H1 holds, because, in this case,

$$T_{B_m} \widehat{\Phi}_{n_m}^{(m)} = \widehat{\Phi}_{n_m}^{(m)}. \tag{3.38}$$

[5]For instance, assume without loss of generality that $\delta_m \in [0, 1/2)$ for all m and take $c = \sqrt{71}$.

[6]For comparison, Section 3.4 deals with i.i.d. considerations in the strong ($\mathbb{P}-$a.s.) sense that follow from the arguments and results in Section 3.3.4.

Thus, (3.36) is proved. Next, note that for every m,

$$0 \leq \inf_{f\in\mathcal{F}_m} \frac{1}{n_m}\sum_{k=1}^{n_m}\int_{\mathbb{R}^{d_m}} |f(x)-\Phi_{m,k}(x)|^2 \mathbb{P}_{X_{m,k}}(\mathrm{d}x)$$
$$\leq \sup_{f\in\mathcal{F}_m} \frac{1}{n_m}\sum_{k=1}^{n_m}\int_{\mathbb{R}^{d_m}} |f(x)-\Phi_{m,k}(x)|^2 \mathbb{P}_{X_{m,k}}(\mathrm{d}x) \leq 2B_m^2, \tag{3.39}$$

and that we can assume, without loss of generality, that $\liminf_m B_m > 0$ (as over sequences m_k with $B_{m_k} \to_k 0$ the conclusion is immediate from (3.39) and (3.38)). Now take

$$\delta_m := \left(\frac{V_{\mathcal{F}_m}\log n_m}{n_m}\right)^{1/2}$$

to verify H1 using (3.37) and (3.39). □

3.3.4 Strong Consistency (Theorem 3.10)

It is possible to profit from the analysis used in the proof of Theorem 3.1 to look for conditions which are linked to the satisfaction of strong consistency for least-squares regression estimates. To illustrate this claim, we depart from the following theorem.

Theorem 3.10. *Assume S1, and consider the random variables*

$$\Delta_m := \inf_{f\in\mathcal{F}_m}\left\{\frac{1}{n_m}\sum_{k=1}^{n_m}(|f(X_{m,k})-Y_{m,k}|^2-|\Phi_{m,k}(X_{m,k})-Y_{m,k}|^2)\right\}$$
$$-\inf_{f\in\mathcal{F}_m}\left\{\frac{1}{n_m}\sum_{k=1}^{n_m}\int_{\mathbb{R}^{d_m}\times\mathbb{R}}(|f(x)-y|^2-|\Phi_{m,k}(x)-y|^2)\mathbb{P}_{X_{m,k},Y_{m,k}}(\mathrm{d}x,\mathrm{d}y)\right\}$$
$$=\inf_{f\in\mathcal{F}_m}\left\{\frac{1}{n_m}\sum_{k=1}^{n_m}(|f(X_{m,k})-Y_{m,k}|^2-|\Phi_{m,k}(X_{m,k})-Y_{m,k}|^2)\right\}$$
$$-\inf_{f\in\mathcal{F}_m}\left\{\frac{1}{n_m}\sum_{k=1}^{n_m}\int_{\mathbb{R}^{d_m}}|f(x)-\Phi_{m,k}(x)|^2\mathbb{P}_{X_{m,k}}(\mathrm{d}x)\right\}. \tag{3.40}$$

If $(n_m^{-\alpha V_{\mathcal{F}_m}})_m$ is summable for some $\alpha>0$,[7] *then for every positive and bounded sequence $(\delta_m)_m$, there exists a constant $C_{(3.41)}>0$ such that*

$$\sup_{f\in\mathcal{F}_m}\frac{1}{n_m}\sum_{k=1}^{n_m}\int_{\mathbb{R}^d}(|T_{B_m}\widehat{\Phi}_{n_m}^{(m)}(x)-\Phi_{m,k}(x)|^2-|f(x)-\Phi_{m,k}(x)|^2)\mathbb{P}_{X_{m,k}}(\mathrm{d}x) \tag{3.41}$$
$$\leq C_{(3.41)}\left(\frac{V_{\mathcal{F}_m}B_m^2\log n_m}{n_m\delta_m}\right)+\delta_m\inf_{f\in\mathcal{F}_m}\left\{\frac{1}{n_m}\sum_{k=1}^{n_m}\int_{\mathbb{R}^{d_m}}|f(x)-\Phi_{m,k}(x)|^2\mathbb{P}_{X_{m,k}}(\mathrm{d}x)\right\}+(1+\delta_m)\Delta_m,$$

except for finitely many m's.
If, in particular, H1 holds and $\Delta_m \to_m 0$, $\mathbb{P}-a.s.$, then

$$\limsup_m \sup_{f\in\mathcal{F}_m}\frac{1}{n_m}\sum_{k=1}^{n_m}\int_{\mathbb{R}^d}(|T_{B_m}\widehat{\Phi}_{n_m}^{(m)}(x)-\Phi_{m,k}(x)|^2-|f(x)-\Phi_{m,k}(x)|^2)\mathbb{P}_{X_{m,k}}(\mathrm{d}x)\leq 0,$$

$\mathbb{P}-a.s.$

[7] For instance if $n_m \geq m^\gamma$ for some $\gamma>0$ (take $\alpha = 1+\frac{1}{\gamma}$).

Proof. For the proof, see Section 3.5. See also Remark 3.19 for considerations in the case of infinite complexity. □

Remark 3.11. Note that, if H2 holds, we can take δ_m as in (3.31).

The following is an immediate consequence of Theorem 3.10.

Corollary 3.12. *Assume H1, assume that $(n_m^{-\alpha V_{\mathcal{F}_m}})_m$ is summable for some $\alpha > 0$, and assume that the uniform strong law of large numbers*

$$\lim_m \sup_{f\in\mathcal{F}_m} \frac{1}{n_m}\sum_{k=1}^{n_m}\Big(|f(X_{m,k})-Y_{m,k}|^2-|\Phi_{m,k}(X_{m,k})-Y_{m,k}|^2 - \int_{\mathbb{R}^d}(|f(x)-y|^2-|\Phi_{m,k}(x)-y|^2)\mathbb{P}_{X_{m,k},Y_{m,k}}(\mathrm{d}x,\mathrm{d}y)\Big)=0, \tag{3.42}$$

holds $\mathbb{P}-a.s.$. Then, with probability one,

$$\limsup_m \sup_{f\in\mathcal{F}_m}\frac{1}{n_m}\sum_{k=1}^{n_m}\int_{\mathbb{R}^d}(|T_{B_m}\widehat{\Phi}_{n_m}^{(m)}(x)-\Phi_{m,k}(x)|^2-|f(x)-\Phi_{m,k}(x)|^2)\mathbb{P}_{X_{m,k}}(\mathrm{d}x)\le 0. \tag{3.43}$$

As an instance of Corollary 3.12, and as a way of illustrating the advantages of knowing not only (3.34) but (3.41), we get the following strong–consistency version of (3.22).

Corollary 3.13. *Let $\mathcal{F}$ be a family of functions $\mathbb{R}^d\to[-B,B]$, assume that $V_{\mathcal{F}}<\infty$, let $((X_k,Y_k))_k$ be any independent sequence in $\mathbb{R}^d\times\mathbb{R}$ with $\sup_k\{||Y_k||_{\mathbb{P},\infty}\}\le B$ and define, for every m, $\widehat{\Phi}_m$ by (1.2) (with $n=m$). Then, with probability one*

$$\lim_m \sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{k=1}^{m}\int_{\mathbb{R}^d}(|\widehat{\Phi}_m(x)-\mathbb{E}\,[Y_k|X_k=x]\,|^2-|f(x)-\mathbb{E}\,[Y_k|X_k=x]\,|^2)\mathbb{P}_{X_k}(\mathrm{d}x)=0. \tag{3.44}$$

Proof. We apply Corollary 3.12 with $n_m=m$, $\mathcal{F}_m=\mathcal{F}$, and $B_m=B$ for all m. Note again that $T_B\widehat{\Phi}_m=\widehat{\Phi}_m$, that $T_BY=Y$, and that the $\liminf_m$ of the expressions inside the $\lim_m$ in (3.44) is nonnegative because each of its terms is not negative.

The proof of (3.42) in these circumstances is omitted by reasons of space. It can be done by extending the proof of [GKKW02, Theorem 9.1] to the independent, non identically distributed case with the same techniques used to prove the inequalities in Theorems 2.2 and 2.3. An alternative approach is to use [Pol90, Theorem 8.3] by considering, in their notation, $T\equiv\mathcal{F}$ and, for $t\in\mathcal{F}$, $f_i(\omega,t)\equiv t(X_i(\omega))$. □

3.4 A non Glivenko-Cantelli Theorem for i.i.d. Samplings (Theorem 3.14)

In the classical context of i.i.d. sampling $((X_k,Y_k))_{k\in\mathbb{N}}$, $(X_k,Y_k)\sim(X,Y)$, and assuming that we are under S1 with $\mathcal{F}_m=\mathcal{F}$, $B_m=B$, $n_m=m$, and $(X_{m,k},Y_{m,k}):=(X_k,Y_k)$ for all (m,k), the property (3.42) reads as

$$\lim_m\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{k=1}^{m}(g_f(X_k,Y_k)-\mathbb{E}\,[g_f(X,Y)])=0,\quad \mathbb{P}-a.s.,$$

where $g_f(x,y):=|f(x)-y|^2-|\Phi(x)-y|^2$ and $\Phi(x)=\mathbb{E}\,[Y|X=x]$, a property known as the *(strong) Glivenko-Cantelli Property (GCP) for the family $\{g_f: f\in\mathcal{F}\}$ and the measure $\mathbb{P}_{(X,Y)}$*.

The satisfaction of the GCP is itself a topic of intensive research in Empirical Process Theory (see [VW96]), and the message of Corollary 3.12 is that the GCP is indeed sufficient, under the respective hypotheses, to guarantee the strong consistency in the sense of (3.43) of a least-squares regression scheme.

But there are situations in which we can still verify (3.43) without appealing to the GCP: note indeed that what matters is the convergence $\mathbb{P}-$a.s. to 0, as $m \to \infty$, of the random variables Δ_m in (3.40).

As we shall see in the proof of the following theorem (Section 3.6), this idea can be used to produce a strongly convergent least–squares regression scheme under minimal hypotheses in the i.i.d. case. Notice indeed that, via the scheme in Theorem 3.14, the statistical problem of strong consistency for i.i.d. samples is reduced to an analytic one: it is *always* possible to choose sample sizes and truncation bounds (see condition (3.45)) that guarantee the consistency of a (properly truncated) least–squares regression scheme over increasing families of functions (with controlled complexity) *provided* that such families approximate *some* version of the conditional expectation sought for. This is true *even if* the GCP does *not* hold for the families in consideration.

For an easy application of this principle see Remark 3.17 below.

Theorem 3.14 (A strongly consistent least–squares regression scheme for the i.i.d. case.)**.** *Let $(X_k, Y_k) \sim (X, Y)$ $(k \in \mathbb{N})$ be i.i.d. random variables in $\mathbb{R}^d \times \mathbb{R}$ with $Y \in L^2_{\mathbb{P}}$ and associated regression function $\Phi(x) = \mathbb{E}[Y|X = x]$. Let $(\mathcal{F}_m)_m$ be an increasing sequence of families of functions with $V_{\mathcal{F}_m} < \infty$ for all m, let $(B_m)_{m\geq 1}$ be a nondecreasing sequence with $B_m \to_m \infty$, assume that $(n_m)_m$ is such that $(n_m^{-\alpha V_{\mathcal{F}_m}})_m$ is summable for some $\alpha > 0$, assume that*

$$\frac{V_{\mathcal{F}_m} B_m^2 \log n_m}{n_m} \to_m 0, \tag{3.45}$$

and let $\widehat{\Phi}^{(m)}_{n_m}$ be given by (1.2) with $\mathcal{F}$ replaced by $\mathcal{F}_m$ and $(X_k, Y_k)_{k=1}^n$ replaced by $(X_k, T_{B_m} Y_k)_{k=1}^{n_m}$. Then, with probability one,

$$\limsup_m \int_{\mathbb{R}^d} |T_{B_m} \widehat{\Phi}^{(m)}_{n_m}(x) - \Phi(x)|^2 \mathbb{P}_X(dx) \leq \lim_m \inf_{f \in \mathcal{F}_m} \int_{\mathbb{R}^d} |f(x) - \Phi(x)|^2 \mathbb{P}_X(dx). \tag{3.46}$$

If, in particular, Φ is a limit point in $L^2_{\mathbb{P}_X}$ of $\mathcal{F}_\infty := \bigcup_m \mathcal{F}_m$, then the assumption (3.45) implies that $T_{B_m} \widehat{\Phi}^{(m)}_{n_m} \to_m \Phi$ in $L^2_{\mathbb{P}_X}$, with probability one.

Proof. See Section 3.6. □

Remark 3.15 (The case of bounded response)**.** If Y is essentially bounded ($||Y||_{\mathbb{P},\infty} < \infty$) one can take $B_m := ||Y||_{\mathbb{P},\infty}$ in the truncations for every m and the conclusion (3.46) holds true if

$$\frac{V_{\mathcal{F}_m} \log n_m}{n_m} \to_m 0.$$

Remark 3.16 (Possible presence of nonintegrable functions)**.** We stress out also the fact that the assumption $\mathcal{F}_m \subset L^2_{\mathbb{P}_X}$ is *not* needed in Theorem 3.14 (in particular, we proved strong consistency under the assumption that $\mathbb{E}[Y|X = \cdot]$ can be approximated in $L^2_{\mathbb{P}_X}$ by elements of $\cup_m \mathcal{F}_m \cap L^2_{\mathbb{P}_X}$, without requiring the full embedding $\cup_m \mathcal{F}_m \subset L^2_{\mathbb{P}_X}$). For an example illustrating how this simplifies, take $X \in L^4_{\mathbb{P}} \setminus L^5_{\mathbb{P}}$ real valued, $Y = X^2 \in L^2_{\mathbb{P}}$, and $\mathcal{F}_m \equiv \mathcal{F}$ as the space of polynomial functions of degree at most 3 (and $X^3 \notin L^2_{\mathbb{P}}$).

Remark 3.17 (Universally consistent schemes)**.** To further stress out the unifying nature of Theorem 3.14, let us give a list of cases for which proving the strong consistency

$$T_{B_m}\widehat{\Phi}^{(m)}_{n_m} \to_m \Phi, \qquad \mathbb{P}-a.s.$$

is easy or simplified via this result (always under (3.45) and, of course, under the assumption $Y \in L^2_{\mathbb{P}}$):

- (Polynomials over a compact set with increasing degrees). When $|X| \in L^\infty_{\mathbb{P}}$ and $\mathcal{F}_m$ is the family of polynomials in d variables of degree at most m (the space generated by $x \mapsto \prod_{j=1}^d x_j^{\alpha_j}$ with $\sup_{1\le j\le d}\alpha_j \le m$), by an application of the Stone–Weierstrass theorem.
- (Piecewise polynomials of fixed maximal degree over increasing partitions). When $|X| \in L^\infty_{\mathbb{P}}$ and $\mathcal{F}_m$ is the space of piecewise polynomial functions in d variables of degree at most M (fixed for all m) over a partition $\mathcal{P}_m$ of a rectangle containing the support of X, where $\mathcal{P}_m$ is refined by $\mathcal{P}_{m+1}$ and the mesh of $\mathcal{P}_m$ goes to zero as $m \to \infty$, by the fact that the set of simple functions in $\mathcal{P}_m$ is asymptotically dense in $L^2_{\mathbb{P}_X}$. Compare with [GKKW02, Theorem 10.4].
- (Neural networks). When $\mathcal{F}_m$ is the class of neural networks defined on [GKKW02, Equation (16.2)], by [GKKW02, Lemma 16.1 and Theorem A.1]. Compare with [GKKW02, Theorem 16.1].
- (RBF networks). When $\mathcal{F}_m$ is a radial basis function network (RBF network) as in [GKKW02, Equation (17.4)], by [GKKW02, Lemma 17.1]. Compare with [GKKW02, Theorem 17.1].

Remark 3.18 (Independence of Theorem 3.14 from the Glivenko-Cantelli property)**.** Considering the first case (polynomial functions) in Remark 3.17, and taking $(X,Y) := (X,X^2)$ for some bounded, non-constant real random variable X we see that, since $f_a(x) := ax^2 \in \mathcal{F}_m$ for all $a \in \mathbb{R}$ and all $m \ge 2$, the GCP (3.42) does *not* hold for any i.i.d. copy $((X_k,Y_k))_k$ of (X,Y). This is of course true whenever X is not constant and $\mathcal{F}_m$ is a vector space and it shows that, while sufficient, (3.42) is not necessary for the consistency (3.46) established in Theorem 3.14.

3.5 Proof of Theorem 3.10

We will use in this case the notation

$$A^{(m)}_{n_m}(\mathrm{d}x) := \frac{1}{n_m}\sum_{k=1}^{n_m}\delta_{X_{m,k}}(\mathrm{d}x), \qquad \mu^{(m)}_{n_m}(\mathrm{d}x) := \frac{1}{n_m}\sum_{k=1}^{n_m}\mathbb{P}_{X_{m,k}}(\mathrm{d}x),$$

[and similarly for $\widetilde{A^{(m)}_{n_m}}$, $\widetilde{\mu^{(m)}_{n_m}}$] for the empirical and the average measures associated to $(X_{m,k})_{k=1}^{n_m}$ [to $(X_{m,k},Y_{m,k})_{k=1}^{n_m}$]. The notation used in the previous proof is extended in a likewise manner. Let $\lambda_m = 1+\delta_m$. Then proceeding as in (3.3) and (3.4) we get

$$\begin{aligned}
&\mu^{(m)}_{n_m}|T_{B_m}\widehat{\mathbf{\Phi}}^{(m)}_{n_m} - \mathbf{\Phi}_{m,1:n_m}|^2 - \inf_{f\in\mathcal{F}_m}\mu^{(m)}_{n_m}|\mathbf{f}-\mathbf{\Phi}_{m,1:n_m}|^2\\
&= \widetilde{\mu^{(m)}_{n_m}}\mathbf{g}^{(m)}_{T_{B_m}\widehat{\Phi}^{(m)}_{n_m}} - \lambda_m\widetilde{A^{(m)}_{n_m}}\mathbf{g}^{(m)}_{T_{B_m}\widehat{\Phi}^{(m)}_{n_m}} + \lambda_m\widetilde{A^{(m)}_{n_m}}\mathbf{g}^{(m)}_{T_{B_m}\widehat{\Phi}^{(m)}_{n_m}} - \inf_{f\in\mathcal{F}_m}\mu^{(m)}_{n_m}|\mathbf{f}-\mathbf{\Phi}_{m,1:n_m}|^2\\
&\le \widetilde{\mu^{(m)}_{n_m}}\mathbf{g}^{(m)}_{T_{B_m}\widehat{\Phi}^{(m)}_{n_m}} - \lambda_m\widetilde{A^{(m)}_{n_m}}\mathbf{g}^{(m)}_{T_{B_m}\widehat{\Phi}^{(m)}_{n_m}} + \lambda_m\Big(\inf_{f\in\mathcal{F}_m}\widetilde{A^{(m)}_{n_m}}\mathbf{g}^{(m)}_f - \inf_{f\in\mathcal{F}_m}\mu^{(m)}_{n_m}|\mathbf{f}-\mathbf{\Phi}_{m,1:n_m}|^2\Big)\\
&\quad + \delta_m\inf_{f\in\mathcal{F}_m}\mu^{(m)}_{n_m}|\mathbf{f}-\mathbf{\Phi}_{m,1:n_m}|^2.
\end{aligned}$$

The proof of (3.41) will be complete once we prove that, always under (3.29),

$$\widetilde{\mu_{n_m}^{(m)}}\mathbf{g}^{(m)}_{T_{B_m}\widehat{\Phi}^{(m)}_{n_m}} - \lambda_m \widetilde{A_{n_m}^{(m)}}\mathbf{g}^{(m)}_{T_{B_m}\widehat{\Phi}^{(m)}_{n_m}} \leq \varepsilon_m,$$

for an appropriate ε_m and except for finitely many m's. We will do this via the Borel-Cantelli lemma.

From now on, we will make use for fixed m of the estimates in the proof of Theorem 3.1 for the inputs $(n_m, B_m, \mathcal{F}_m, \lambda_m)$, where $c = \sqrt{71}$ (so that (3.1) holds for every $\lambda = \lambda_m$[8]). We will proceed letting some constants which do not depend on m (assuming if needed that m is big enough) unspecified in the analysis that follow, in all cases a specification whenever needed would be easy.

For $\lambda = \lambda_m$ and the corresponding input $(n_m, \mathcal{F}_m, \lambda_m)$, let G_0, $G_1^{(m)}$, and $b^{(m)}$ be given by (3.9), and let $t_{n_m}^{(m)}$ be given by (3.10) with $\lambda = \lambda_m$ and $n = n_m$. Notice that, in virtue of (3.16),

$$t_{n_m}^{(m)} \leq \frac{c(c+1)}{4}\frac{\lambda_m^2}{\delta_m n_m}.$$

Let ε_m be given, for some $\gamma \geq 1$ to be specified later, by

$$\varepsilon_m := 4\gamma c(c+1)\lambda_m^2\left(\frac{V_{\mathcal{F}_m}B_m^2\log n_m}{\delta_m n_m}\right) \geq 4c(c+1)\lambda_m^2\frac{B_m^2}{\delta_m n_m} \geq 16B_m^2 t_{n_m}^{(m)}. \tag{3.47}$$

Then by an elementary extension of (3.11) and (3.13) (which are derived for the explicit bound $B = 1/4$) the estimate

$$\begin{aligned}\mathbb{P}&\left(\widetilde{\mu_{n_m}^{(m)}}\mathbf{g}^{(m)}_{T_{B_m}\widehat{\Phi}^{(m)}_{n_m}} - \lambda_m\widetilde{A_{n_m}^{(m)}}\mathbf{g}^{(m)}_{T_{B_m}\widehat{\Phi}^{(m)}_{n_m}} \geq \varepsilon_m\right)\\ &\leq 3G_0\left(\frac{e}{G_1^{(m)}t_{n_m}^{(m)}}\log\frac{3e}{2G_1^{(m)}t_{n_m}^{(m)}}\right)^{V_{\mathcal{F}_m}}\exp\left(-b^{(m)}n_m\frac{\varepsilon_m}{16B_m^2}\right)\end{aligned} \tag{3.48}$$

holds. We deduce further from this that for some positive constants a_1, a_2, a_3, the above probability is bounded by

$$\leq a_1\exp\left(a_2V_{\mathcal{F}_m}(1-\log(G_1^{(m)}t_{n_m}^{(m)})) - a_3\frac{b^{(m)}\varepsilon_m n_m}{B_m^2}\right). \tag{3.49}$$

We use now the nonasymptotic estimates[9] (see (3.9), (3.18) and (3.19))

$$\frac{b^{(m)}\varepsilon_m n_m}{B_m^2} \sim \frac{\varepsilon_m n_m\delta_m}{B_m^2}, \qquad -\log(G_1^{(m)}t_{n_m}^{(m)}) \sim \log n_m$$

to deduce from (3.49) that for some positive constants a_4, a_5, a_6,

$$\begin{aligned}\mathbb{P}\left(\widetilde{\mu_{n_m}^{(m)}}\mathbf{g}^{(m)}_{T_{B_m}\widehat{\Phi}^{(m)}_{n_m}} - \lambda_m\widetilde{A_{n_m}^{(m)}}\mathbf{g}^{(m)}_{T_{B_m}\widehat{\Phi}^{(m)}_{n_m}} \geq \varepsilon_m\right) &\leq a_1\exp\left(a_4V_{\mathcal{F}_m}\log n_m - a_5\frac{\varepsilon_m n_m\delta_m}{B_m^2}\right)\\ &\leq a_1\exp\left(-a_5V_{\mathcal{F}_m}(\frac{\varepsilon_m}{\alpha_m} - a_6)\log n_m\right)\end{aligned}$$

[8]For the case in which $(\delta_m)_m$ is bounded by an arbitrary constant, use the fact that $V_{\mathcal{F}_m}\log n_m \to_m \infty$ to prove that any $c > 0$ works for all but finitely many m's.

[9]For functions f, g with the same domain, $f \sim g$ is a shortcut for the existence of constants $0 < a < b$ with the property that, for every x, $af(x) \leq g(x) \leq bf(x)$.

$$= a_1 \exp\left(-a_5 V_{\mathcal{F}_m}(4c(c+1)\gamma\lambda_m^2 - a_6)\log n_m\right)$$

where

$$\alpha_m := \frac{V_{\mathcal{F}_m} B_m^2 \log n_m}{\delta_m n_m}.$$

By taking $\gamma > 0$ big enough so that

$$a_5(4c(c+1)\gamma - a_6) \geq \alpha,$$

the conclusion follows from the Borel–Cantelli lemma and the summability of $(n_m^{-\alpha V_{\mathcal{F}_m}})$ with $C_{(3.41)} := 4c(c+1)\gamma \sup_m\{\lambda_m^2\}$. □

Remark 3.19 (Complement to Remarks 3.3 and 3.5)**.** When (3.21) holds for all m with $\mathcal{F}$ replaced by $\mathcal{F}_m$ and α replaced by $\alpha_m \in (0,1)$, and assuming without loss of generality that the corresponding $C_{\mathcal{F}_m}$ satisfies $C_{\mathcal{F}_m} \geq 1$ (compare with Remark 3.5), the proof of Theorem 3.10 can be modified to show that (3.41) holds replacing the term indicated in (3.32) by that in (3.33) provided that, for some $\rho > 0$, $(\exp(-\rho n_m^{\alpha_m}))_m$ is a summable sequence. To see this (further details are left to the reader) use, instead of ε_m in (3.47),

$$\varepsilon'_m := 4\gamma c(c+1) C_{\mathcal{F}_m} \lambda_m^2 \frac{B_m^2}{n_m^{1-\alpha}\delta_m},$$

and replace the estimate (3.48) by the corresponding estimate

$$\mathbb{P}\left(\widetilde{\mu_{n_m}^{(m)}}\mathbf{g}^{(m)}_{T_{B_m}\widehat{\Phi}^{(m)}_{n_m}} - \lambda_m \widetilde{A_{n_m}^{(m)}}\mathbf{g}^{(m)}_{T_{B_m}\widehat{\Phi}^{(m)}_{n_m}} \geq \varepsilon'_m\right) \leq 3G_0 \exp\left(C_{\mathcal{F}_m}(G_1^{(m)} t_{n_m}^{(m)})^{-\alpha_m} - b^{(m)} n_m \frac{\varepsilon'_m}{16 B_m^2}\right).$$

The respective consequences for strong consistency can be deduced in an easy manner.

3.6 Proof of Theorem 3.14

First notice that, since $(\mathcal{F}_m)_m$ is increasing, the sequence

$$(\varepsilon_m)_m := \left(\inf_{f\in\mathcal{F}_m} \int_{\mathbb{R}^d} |f(x) - \Phi(x)|^2 \mathbb{P}_X(\mathrm{d}x)\right)_m \tag{3.50}$$

is non-increasing, and note that we can assume that this sequence is ultimately finite, i.e., that $\int_{\mathbb{R}^d} |f(x) - \Phi(x)|^2 \mathbb{P}_X(\mathrm{d}x) < \infty$ for some $f \in \mathcal{F}_\infty := \cup_m \mathcal{F}_m$, as otherwise (3.46) is trivially true.

Let $\Phi^{(L)}$ be specified, for every $L \geq 0$, by

$$\Phi^{(L)}(x) := \mathbb{E}\left[T_L Y | X = x\right].$$

To clarify some explanations later, we also introduce the $\mathbb{P}_X$-square integrable functions Φ_{abs} and $\Phi_{lar(L)}$ $(L > 0)$ characterized $\mathbb{P}_X-$a.s. by the equations

$$\Phi_{abs}(x) = \mathbb{E}\left[|Y||X = x\right], \qquad \Phi_{lar(L)}(x) := \mathbb{E}\left[|Y|\mathbf{1}_{|Y|>L}|X = x\right],$$

and we observe that, by Jensen's inequality and the definition of T_L, for $\mathbb{P}_X-$a.e. x

$$\max\{\sup_m |\Phi^{(B_m)}(x)|, |\Phi(x)|\} \leq \Phi_{abs}(x),$$

$$|\Phi(x) - \Phi^{(L)}(x)| \leq \mathbb{E}\left[|Y - T_L Y||X = x\right] \leq \Phi_{lar(L)}(x),$$

and that if $L_1 \leq L_2$, then $\Phi_{lar(L_2)} \leq \Phi_{lar(L_1)}$, $\mathbb{P}_X-$a.s. Likewise notice that

$$\int_{\mathbb{R}^d} |\Phi(x) - \Phi^{(B_m)}(x)|^2 \mathbb{P}_X(\mathrm{d}x) = \mathbb{E}\left[|\mathbb{E}\left[Y|X\right] - \mathbb{E}\left[T_{B_m} Y|X\right]|^2\right] \leq \mathbb{E}\left[|Y - T_{B_m} Y|^2\right], \tag{3.51}$$

and since $B_m \to_m \infty$, we deduce from (3.51) that $\Phi^{(B_m)} \to_m \Phi$ in $L^2_{\mathbb{P}_X}$. In particular, it follows easily that (3.46) is equivalent to the assertion

$$\limsup_m \left(\int_{\mathbb{R}^d} |T_{B_m} \widehat{\Phi}^{(m)}_{n_m}(x) - \Phi^{(B_m)}(x)|^2 \mathbb{P}_X(\mathrm{d}x) - \inf_{f \in \mathcal{F}_m} \int_{\mathbb{R}^d} |f(x) - \Phi^{(B_m)}(x)|^2 \mathbb{P}_X(\mathrm{d}x) \right) \leq 0, \quad \mathbb{P} - a.s.$$

We apply Theorem 3.10 with $Y_{m,k} = T_{B_m} Y_k$. Note that, with the notation in that theorem,

$$\Phi_{m,k} = \Phi^{(B_m)},$$

$\mathbb{P}_X-$a.s. for every m and every $1 \leq k \leq n_m$, and that we are here under the hypothesis H2. Specializing the notation from the previous proof in an obvious way (in particular $\mu := \mathbb{P}_X = \mu_m$ for all m) notice that, if Δ_m is the random variable in Theorem 3.10 then, again by the convergence of $\Phi^{(B_m)} \to_m \Phi$ in $L^2_{\mathbb{P}_X}$,

$$\begin{aligned}
\limsup_m \Delta_m &:= \limsup_m \left(\inf_{f \in \mathcal{F}_m} \widetilde{A^{(m)}_{n_m}} \mathbf{g}^{(m)}_f - \inf_{f \in \mathcal{F}_m} \mu|f - \Phi^{(B_m)}|^2 \right) \\
&= \limsup_m \left(\inf_{f \in \mathcal{F}_m} \widetilde{A^{(m)}_{n_m}} \mathbf{g}^{(m)}_f - \inf_{f \in \mathcal{F}_m} \mu|f - \Phi|^2 \right).
\end{aligned}$$

It suffices therefore by (3.41) to see that, under the given hypotheses,

$$\limsup_m (\inf_{f \in \mathcal{F}_m} \widetilde{A^{(m)}_{n_m}} \mathbf{g}^{(m)}_f - \inf_{f \in \mathcal{F}_m} \mu|f - \Phi|^2) \leq 0, \quad \mathbb{P} - a.s. \tag{3.52}$$

Proof of (3.52). Let $\varepsilon_\infty = \lim_m \varepsilon_m$ (see (3.50)) and, for any $\delta > 0$ given, let $M_\delta \in \mathbb{N}$ and $f_\delta \in \mathcal{F}_{M_\delta}$ be such that for every $m \geq M_\delta$

$$\varepsilon_m \leq \mu|f_\delta - \Phi|^2 < \varepsilon_\infty + \delta \leq \varepsilon_m + \delta < \infty,$$

note in particular that $f_\delta \in L^2_{\mathbb{P}_X}$. Now, since

$$\inf_{f \in \mathcal{F}_m} \widetilde{A^{(m)}_{n_m}} \mathbf{g}^{(m)}_f - \inf_{f \in \mathcal{F}_m} \mu|f - \Phi|^2 \leq \widetilde{A^{(m)}_{n_m}}(|\mathbf{f}_\delta - T_{B_m} \mathbf{y}_{1:n_m}|^2 - |\mathbf{\Phi}^{(B_m)} - T_{B_m} \mathbf{y}_{1:n_m}|^2) - \mu|f_\delta - \Phi|^2 + \delta$$

for $m \geq M_\delta$, it suffices therefore to prove that

$$\limsup_m \widetilde{A^{(m)}_{n_m}}(|\mathbf{f}_\delta - T_{B_m} \mathbf{y}_{1:n_m}|^2 - |\mathbf{\Phi}^{(B_m)} - T_{B_m} \mathbf{y}_{1:n_m}|^2) \leq \mu|f_\delta - \Phi|^2,$$

$\mathbb{P}-$a.s. Note now that by the Law of Large Numbers

$$\widetilde{A^{(m)}_{n_m}}(|\mathbf{f}_\delta - \mathbf{y}_{1:n_m}|^2 - |\mathbf{\Phi} - \mathbf{y}_{1:n_m}|^2) \to_m \widetilde{\mu}\left(|f_\delta - y|^2 - |\Phi - y|^2\right) = \mu|f_\delta - \Phi|^2,$$

$\mathbb{P}-$a.s., and therefore it suffices to prove that, $\mathbb{P}-$a.s.,

$$\limsup_m \left(\widetilde{A^{(m)}_{n_m}}(|\mathbf{f}_\delta - T_{B_m}\mathbf{y}_{1:n_m}|^2 - |\Phi^{(B_m)} - T_{B_m}\mathbf{y}_{1:n_m}|^2) - \widetilde{A^{(m)}_{n_m}}(|\mathbf{f}_\delta - \mathbf{y}_{1:n_m}|^2 - |\mathbf{\Phi} - \mathbf{y}_{1:n_m}|^2) \right) \leq 0. \tag{3.53}$$

To do so take $L \geq 0$ and consider m large enough to have $m \geq M_\delta$ and $B_m \geq L$, then

$$\begin{aligned}
&\widetilde{A^{(m)}_{n_m}} \left(|\mathbf{f}_\delta - T_{B_m}\mathbf{y}_{1:n_m}|^2 - |\mathbf{\Phi}^{(B_m)} - T_{B_m}\mathbf{y}_{1:n_m}|^2) - \widetilde{A^{(m)}_{n_m}}(|\mathbf{f}_\delta - \mathbf{y}_{1:n_m}|^2 - |\mathbf{\Phi} - \mathbf{y}_{1:n_m}|^2 \right) \\
\leq &\widetilde{A^{(m)}_{n_m}} \left(|T_{B_m}\mathbf{y}_{1:n_m} - \mathbf{y}_{1:n_m}|(2|\mathbf{f}_\delta| + 2|\mathbf{y}_{1:n_m}|) + (|T_{B_m}\mathbf{y}_{1:n_m} - \mathbf{y}_{1:n_m}| + |\mathbf{\Phi}^{(B_m)} - \mathbf{\Phi}|)(|\mathbf{\Phi}^{(B_m)}| + |\mathbf{\Phi}| + 2|\mathbf{y}_{1:n_m}|) \right) \\
\leq &\widetilde{A^{(m)}_{n_m}} \left(|T_L\mathbf{y}_{1:n_m} - \mathbf{y}_{1:n_m}|(|2\mathbf{f}_\delta| + |2\mathbf{y}_{1:n_m}|) + (|T_L\mathbf{y}_{1:n_m} - \mathbf{y}_{1:n_m}| + \mathbf{\Phi}_{lar(L)})(2\mathbf{\Phi}_{abs} + 2|\mathbf{y}_{1:n_m}|) \right),
\end{aligned}$$

and therefore, by the Law of Large Numbers, Hölder's inequality, and Jensen's inequality, there exists a constant C not depending of L such that

$$\begin{aligned}
\limsup_m \widetilde{A^{(m)}_{n_m}} &\left(|\mathbf{f}_\delta - T_{B_m}\mathbf{y}_{1:n_m}|^2 - |\mathbf{\Phi}^{(B_m)} - T_{B_m}\mathbf{y}_{1:n_m}|^2) - \widetilde{A^{(m)}_{n_m}}(|\mathbf{f}_\delta - \mathbf{y}_{1:n_m}|^2 - |\mathbf{\Phi} - \mathbf{y}_{1:n_m}|^2 \right) \\
&\leq C((\mathbb{E}\left[|T_L Y - Y|^2\right])^{1/2} + (\mathbb{E}\left[(\Phi_{lar(L)}(X))^2\right])^{1/2}) \leq 2C(\mathbb{E}\left[|Y|^2 \mathbf{1}_{|Y|>L}\right])^{1/2},
\end{aligned}$$

$\mathbb{P}-$a.s. From here, (3.53) follows by letting $L \to \infty$. □

References

[Ben62] G. Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, March 1962.

[Ber24] S.N. Bernstein. On a modification of Chebyshev's inequality and of the error formula of Laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math.*, 1, 1924.

[Can33] F. P. Cantelli. Sulla determinazione empirica delle leggi di probabilità. *Giorn. Ist. Ital. Attuari*, 4:421–424, 1933.

[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.

[DG95] D. Duffie and P.W. Glynn. Efficient Monte-Carlo simulation of security prices. *Ann. Appl. Probab.*, 5(4):897–905, 1995.

[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.

[Gil08] M.B. Giles. Multilevel Monte-Carlo path simulation. *Operation Research*, 56:607–617, 2008.

[Gil15] M.B. Giles. Multilevel Monte Carlo methods. *Acta Numerica*, 24:259–328, 2015.

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

[Gli33] V. Glivenko. Sulla determinazione empirica delle leggi di probabilità. *Giorn. Ist. Ital. Attuari*, 4:92–99, 1933.

[GW96] S. van de Geer and M. Wegkamp. Consistency for the least squares estimator in nonparametric regression. *The Annals of Statistics*, 24(6):2513–2523, 1996.

[Hei01] S. Heinrich. Multilevel Monte-Carlo Methods. In *LSSC '01 Proceedings of the Third International Conference on Large-Scale Scientific Computing*, volume 2179 of *Lecture Notes in Computer Science*, pages 58–67. Springer-Verlag, 2001.

[Hoe63] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[HTF09] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning. Data mining, inference, and prediction.* Springer Series in Statistics. New York, NY: Springer, second edition, 2009.

[HW17] Q. Han and Jon A. Wellner. A sharp multiplier inequality with applications to heavy-tailed regression problems. *arXiv:1706.02410v1*, 2017.

[Kah68] J.-P. Kahane. *Some random series of functions.* Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, second edition, 1968.

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

[Kos03] M.R. Kosorok. Bootstraps of sums of independent but not identically distributed stochastic processes. *Journal of Multivariate Analysis*, 84:299–318, 2003.

[KP10] P.E. Kloeden and E. Platen. *Numerical solution of stochastic differential equations.* Applications of Mathematics 23. Berlin: Springer, fourth edition, 2010.

[MM12] M. Mohri and A. Muñoz Medina. New analysis and algorithm for learning with drifiting distributions. In *Algorithmic Learning Theory. 23rd International Conference, ALT 2012, Lyon, France, October 29–31, 2012. Proceedings.*, 2012.

[Oks00] B. Oksendal. *Stochastic Differential Equations, an introduction with applications.* Springer-Verlag Heidelberg New York, fifth edition, 2000.

[PG99] V.H. de la Peña and E. Giné. *Decoupling. From Dependence to Indepedence.* Probability and its Applications. Springer, 1999.

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

[Vap00] V.N. Vapnik. *The Nature of Statistical Learning Theory.* Statistics for Engineering and Information Science. Springer, second edition, 2000.

[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.

[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.

[Wel81] J.A. Wellner. A Glivenko-Cantelli theorem for empirical measures of independent but non-identically distributed random variables. *Stochastic Processes and their Applications*, 11(3):309–312, 1981.

[Zui78] M.C.A. van Zuijlen. Properties of the empirical distribution function for independent non-identically distributed random variables. *The Annals of Probability*, 6(2):250–266, 1978.