ESTIMATING A REGRESSION FUNCTION IN EXPONENTIAL FAMILIES BY MODEL SELECTION

JUNTONG CHEN

ABSTRACT. Let $X_1 = (W_1, Y_1), \dots, X_n = (W_n, Y_n)$ be n pairs of independent random variables. We assume that, for each $i \in \{1, \dots, n\}$, the conditional distribution of Y_i given W_i belongs to a one-parameter exponential family with parameter $\boldsymbol{\gamma}^\star(W_i) \in \mathbb{R}$, or at least, is close enough to a distribution of this form. The objective of the present paper is to estimate these conditional distributions on the basis of the observation $\boldsymbol{X} = (X_1, \dots, X_n)$ and to do so, we propose a model selection procedure together with a non-asymptotic risk bound for the resulted estimator with respect to a Hellinger-type distance. When $\boldsymbol{\gamma}^\star$ does exist, the procedure allows to obtain an estimator $\widehat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}^\star$ adapted to a wide range of the anisotropic Besov spaces. When $\boldsymbol{\gamma}^\star$ has a general additive or multiple index structure, we construct suitable models and show the resulted estimators by our procedure based on such models can circumvent the curse of dimensionality. Moreover, we consider model selection problems for ReLU neural networks and provide an example where estimation based on neural networks enjoys a much faster converge rate than the classical models. Finally, we apply this procedure to solve variable selection problem in exponential families. The proofs in the paper rely on bounding the VC dimensions of several collections of functions, which can be of independent interest.

1. INTRODUCTION

We observe n pairs of independent random variables $X_i = (W_i, Y_i)$, for $i \in \{1, \dots, n\}$ with values in a measurable product space $(\mathscr{W} \times \mathscr{Y}, \mathcal{W} \otimes \mathcal{Y})$ and assume (even if this may not be true) that there exists an unknown function $\boldsymbol{\gamma}^\star$ on $\mathscr{W}$ such that for each $i \in \{1, \dots, n\}$, the conditional distribution of Y_i given W_i belongs to a one parameter exponential family with parameter $\boldsymbol{\gamma}^\star(W_i) \in \mathbb{R}$. From the observation $\boldsymbol{X} = (X_1, \dots, X_n)$, our aim is to estimate the conditional distributions of Y_i given W_i. Our approach is based on an estimation of the potential $\boldsymbol{\gamma}^\star$ (may not exist). When such

Date: March 13, 2022.

2020 *Mathematics Subject Classification.* Primary 62G05, 62G35; Secondary 62J12.

Key words and phrases. Model selection, adaptive estimation, generalized additive models, multiple index models, ReLU neural networks, variable selection.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 811017.

a $\boldsymbol{\gamma}^\star$ does exist, the above statistical setting includes as particular cases those of binary, Gaussian and Poisson regressions as well as exponential multiplicative regression.

We are not aware of many results in the literature that tackle these regression problems and establish a risk bound for the proposed estimator $\widehat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}^\star$. When the exponential family is given under its canonical form and $\mathscr{W} = [0,1]$, Kolaczyk and Nowak (2005) proposed a piecewise polynomial estimator $\widehat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}^\star$ under the assumption that the natural parameter is a smooth function of the mean. They showed that their estimator achieves the usual rate $n^{-2\alpha/(1+2\alpha)}$ over the Besov spaces with regularity $\alpha > 0$ up to a logarithmic factor with respect to the squared Hellinger loss. When the family is a univariate natural exponential family where the variances of the distributions are quadratic functions of the means, Antoniadis et al. (2001) proposed a wavelet shrinkage method to estimate $\boldsymbol{\gamma}^\star$ while Brown et al. (2010) parametrized the exponential family by its mean and transformed the problem of estimating $\boldsymbol{\gamma}^\star$ into homoscedastic Gaussian regression estimation by stabilizing the variance. Under the assumption that $\boldsymbol{\gamma}^\star$ belongs to the Besov spaces and is bounded from both below and above, Brown et al. (2010) showed their estimator achieves the converge rate $n^{-2\alpha/(1+2\alpha)}$ with respect to the squared $\mathbb{L}_2$-loss. All of the literatures mentioned above make strong assumptions on the distributions of the covariates W_i where they require them to be known (for example deterministic) or partly known. Moreover, none of them consider the problem of estimation under a possible misspecification framework.

Recently, Baraud and Chen (2020) proposed a robust procedure based on the ρ-estimation to estimate $\boldsymbol{\gamma}^\star$. Their approach is restricted to the case of a single model. Up to a numerical constant, the risk of their estimator $\widehat{\boldsymbol{\gamma}}$ on the model is bounded by the sum of an approximation term and a complexity term. Such an estimation procedure is satisfactory if we know in advance a suitable model for $\boldsymbol{\gamma}^\star$, i.e. a model which is not too complex and provides a good enough approximation of $\boldsymbol{\gamma}^\star$. However, such a model may not be easy to design without any prior information and a safer approach is to consider a family of candidate models instead and let the data decide which is the most appropriate one for estimating $\boldsymbol{\gamma}^\star$.

1.1. **Our contributions.** In this paper, we consider estimating the conditional distributions $R_i^\star(W_i)$ of Y_i given W_i by model selection. Our main contributions are as follows.

(i) We propose a model selection procedure to estimate the conditional distributions and establish a non-asymptotic risk bound for the resulted estimator. Our approach is based on the presumption that there exists an unknown $\boldsymbol{\gamma}^\star$ on $\mathscr{W}$ belonging to some of our models such that $R_i^\star(W_i)$ is of the form $R_{\boldsymbol{\gamma}^\star(W_i)}$ for all $i \in \{1,\dots,n\}$.

However, our approach is not restricted to this assumption. This is to say, we allow our statistical models to be slightly misspecified: $R_i^\star(W_i)$ may not be exactly of the form $R_{\boldsymbol{\gamma}^\star(W_i)}$ and even if they were, $\boldsymbol{\gamma}^\star$ may not belong to any of our models. What we really assume is the form $R_{\boldsymbol{\gamma}(W_i)}$ with some $\boldsymbol{\gamma}$ belonging to our models provides a suitable approximation of the conditional distributions.

(ii) When $X_1, \ldots, X_n$ are i.i.d., this model selection procedure solves adaptation and variable selection problems in exponential families.

(iii) In i.i.d. case, when the dimensionality d of covariate W is large, the converge rate of estimating $\boldsymbol{\gamma}^\star$ can be extremely slow which is, as a well-known phenomenon, called the curse of dimensionality. When $\boldsymbol{\gamma}^\star$ has some particular structures or at least close to some function with such features, we consider model selection problems based on the composite piecewise polynomials and ReLU neural networks and show that the resulted estimators by our procedure based on such models can circumvent the curse of dimensionality. The structures discussed in the paper includes genaralized additive structure, multiple index structure and multiple composition structure.

(iv) In i.i.d. case, when $\boldsymbol{\gamma}^\star$ belongs to the Takagi class we provide an example where estimation based on ReLU neural networks results in an estimator converging to $\boldsymbol{\gamma}^\star$ with parametric rate although $\boldsymbol{\gamma}^\star$ has very little smoothness. At least for such an example, neural networks outperform all the other traditional approximation methods, e.g. piecewise polynomials and wavelets.

(v) We construct models to approximate general additive and multiple index functions and derive VC dimension bounds for them. Besides, we adapt the VC dimension result of ReLU neural networks to the sparse setting. These results can be of independent interest for readers.

The paper is organized as follows. We introduce the specific statistical framework and set notations in Section 2. An estimation procedure based on model selection is proposed in Section 3 together with the non-asymptotic exponential deviation inequalities. We then discuss the adaptive estimation problem in exponential families when the regression function belongs to anisotropic Besov spaces as an application in Section 4. We show that under a suitable parametrization of exponential families, our estimator is adaptive over a wide range of the anisotropic Besov spaces with the risk bound independent of choice of the exponential family. In Section 5, we consider the applications of our procedure to two examples of the structural assumptions, general additive functions and multiple index functions, to circumvent the curse of dimensionality. Estimation by model selection based on ReLU neural networks is discussed in Section 6 and variable selection problem in generalized linear models is considered in Section 7. Finally, all the proofs of this paper can be found in the appendix.

2. The statistical setting

As already mentioned, we observe n independent pairs of random variables $X_1 = (W_1, Y_1), \dots, X_n = (W_n, Y_n)$ with values in a measurable product space $(\mathscr{X}, \mathcal{X}) = (\mathscr{W} \times \mathscr{Y}, \mathcal{W} \otimes \mathcal{Y})$. We assume that for each $i \in \{1, \dots, n\}$, the conditional distribution of Y_i given W_i is given by $R_i^\star(W_i)$, where $R_i^\star$ is a measurable function from $(\mathscr{W}, \mathcal{W})$ to the set of all probabilities on $(\mathscr{Y}, \mathcal{Y})$ equipped with the Borel σ-algebra $\mathscr{T}$ associated to the Hellinger distance. We denote $\mathbf{R}^\star$ by n-tuple $(R_1^\star, \dots, R_n^\star)$.

Let I be a non-trivial interval of $\mathbb{R}$, i.e. the interior of I is not empty and $\widetilde{\mathscr{Q}} = \{R_\gamma = \overline{r}_\gamma \cdot \mu,\ \gamma \in I\}$ be an exponential family under its general form on the measured space $(\mathscr{Y}, \mathcal{Y}, \mu)$. More precisely, $\widetilde{\mathscr{Q}}$ is a family of probabilities on $(\mathscr{Y}, \mathcal{Y})$ admitting densities $\overline{r}_\gamma$ with respect to μ of the form, for all $y \in \mathscr{Y}$ and $\gamma \in I$

$$\overline{r}_\gamma(y) = e^{u(\gamma)T(y) - B(\gamma)} a(y) \text{ where } B(\gamma) = \log\left[\int_{\mathscr{Y}} e^{u(\gamma)T(y)} a(y) d\mu(y)\right], \tag{1}$$

T is a real-valued measurable function on $(\mathscr{Y}, \mathcal{Y})$ which does not coincide with a constant $\nu = a \cdot \mu$-a.e., u is a continuous, strictly monotone function on I and a is a nonnegative function on $\mathscr{Y}$. For convenience, we denote

$$r_\gamma(y) = e^{u(\gamma)T(y) - B(\gamma)}, \quad \text{for all } y \in \mathscr{Y} \text{ and } \gamma \in I, \tag{2}$$

and rewrite $\widetilde{\mathscr{Q}} = \{R_\gamma = r_\gamma \cdot \nu,\ \gamma \in I\}$.

Our estimator takes the form of a mapping $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}} : \mathbf{w} = (\boldsymbol{w}_1, \dots, \boldsymbol{w}_n) \in \mathscr{W}^n \mapsto (R_{\widehat{\boldsymbol{\gamma}}(\boldsymbol{w}_1)}, \dots, R_{\widehat{\boldsymbol{\gamma}}(\boldsymbol{w}_n)})$ with values in $\widetilde{\mathscr{Q}}^n$, where $\widehat{\boldsymbol{\gamma}}$ is a (random) function from $\mathscr{W}$ into I. When the mapping $\mathbf{R}^\star$ is also of this form, i.e. $\mathbf{R}^\star = \mathbf{R}_{\boldsymbol{\gamma}^\star}$ for some (deterministic) function $\boldsymbol{\gamma}^\star : \mathscr{W} \to I$, $\widehat{\boldsymbol{\gamma}}$ provides an estimator of the so called *regression function*.

In order to evaluate the performance of our estimator $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}}$ of $\mathbf{R}^\star$, we introduce a loss function based on the Hellinger distance. Given P and Q two probabilities dominated by some reference measure μ on a measurable space $(A, \mathcal{A})$, the Hellinger distance between $P = p \cdot \mu$ and $Q = q \cdot \mu$ is given by

$$h(P, Q) = \left[\frac{1}{2} \int_A \left(\sqrt{p} - \sqrt{q}\right)^2 d\mu\right]^{1/2}. \tag{3}$$

We remark that (3) does not depend on the choice of μ. We denote $\mathscr{Q}_{\mathscr{W}}$ as the set of all measurable mappings from $(\mathscr{W}, \mathcal{W})$ to the space of probabilities on $(\mathscr{Y}, \mathcal{Y})$ equipped with the topology $\mathscr{T}$ and define $\boldsymbol{\mathscr{Q}}_{\mathscr{W}} = \mathscr{Q}_{\mathscr{W}}^n$. Therefore, both $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}}$ and $\mathbf{R}^\star = (R_1^\star, \dots, R_n^\star)$ belong to $\boldsymbol{\mathscr{Q}}_{\mathscr{W}}$. We endow the space $\boldsymbol{\mathscr{Q}}_{\mathscr{W}}$ with the pseudo Hellinger distance $\mathbf{h}$ defined for $\mathbf{R} = (R_1, \dots, R_n)$ and

$\mathbf{R}' = (R'_1, \ldots, R'_n)$ in $\boldsymbol{\mathscr{Q}}_{\mathscr{W}}$ by

$$\mathbf{h}^2(\mathbf{R}, \mathbf{R}') = \mathbb{E}\left[\sum_{i=1}^{n} h^2\left(R_i(W_i), R'_i(W_i)\right)\right] = \sum_{i=1}^{n} \int_{\mathscr{W}} h^2\left(R_i(w), R'_i(w)\right) dP_{W_i}(w), \tag{4}$$

where h is Hellinger distance defined in (3). We evaluate the performance of the estimator $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}}$ of $\mathbf{R}^\star$ by the quantity $\mathbf{h}^2(\mathbf{R}^\star, \mathbf{R}_{\widehat{\boldsymbol{\gamma}}})$. In the favourable situation where $\boldsymbol{\gamma}^\star$ does exist, we automatically deduce a performance of $\widehat{\boldsymbol{\gamma}}$ with respect to $\boldsymbol{\gamma}^\star$ by the distance $d(\boldsymbol{\gamma}^\star, \widehat{\boldsymbol{\gamma}}) = \mathbf{h}(\mathbf{R}_{\boldsymbol{\gamma}^\star}, \mathbf{R}_{\widehat{\boldsymbol{\gamma}}})$.

When W_i are i.i.d. with the common distribution P_W and $R_i^\star = R^\star$ for all $i \in \{1, \ldots, n\}$, we slightly abuse the notation $h^2(R^\star, R_{\widehat{\boldsymbol{\gamma}}})$ to measure the distance between $R^\star$ and $R_{\widehat{\boldsymbol{\gamma}}}$ which is defined as

$$h^2(R^\star, R_{\widehat{\boldsymbol{\gamma}}}) = \frac{1}{n}\mathbf{h}^2(\mathbf{R}^\star, \mathbf{R}_{\widehat{\boldsymbol{\gamma}}}) = \int_{\mathscr{W}} h^2(R^\star(w), R_{\widehat{\boldsymbol{\gamma}}(w)}) dP_W(w).$$

We end this section by introducing some notations for later use. We denote $\mathbb{N}^*$ the set of all positive natural numbers, $\mathbb{R}_+$ the set of all non-negative real numbers and $\mathbb{R}_+^*$ the set of all positive real numbers. For a set m, we use $|m|$ to denote its cardinality. By $(x)_+$, we mean the function $\max\{0, x\}$. We denote $x \vee y$ the largest value among $\{x, y\}$ while $x \wedge y$ is the smallest. We use the notation $\lfloor x \rfloor$ for any $x \in \mathbb{R}$ to denote the largest integer strictly smaller than x. For a $\mathbf{R} \in \boldsymbol{\mathscr{Q}}_{\mathscr{W}}$ and a set $\mathbf{A} \subset \boldsymbol{\mathscr{Q}}_{\mathscr{W}}$, we define $\mathbf{h}^2(\mathbf{R}, \mathbf{A}) = \inf_{\mathbf{R}' \in \mathbf{A}} \mathbf{h}^2(\mathbf{R}, \mathbf{R}')$. Unless otherwise specified, log denotes the logarithm function with base e. Let $(A, \mathcal{A})$ be a measurable space and μ be a σ-finite measure on $(A, \mathcal{A})$. For $k \in [1, +\infty]$, we define $\mathcal{L}_k(A, \mu)$ the collection of all the measurable functions f on $(A, \mathcal{A}, \mu)$ such that $\|f\|_{k,\mu} < +\infty$, where

$$\|f\|_{k,\mu} = \left(\int_A |f|^k d\mu\right)^{\frac{1}{k}}, \quad \text{for } k \in [1, +\infty),$$

$$\|f\|_{\infty,\mu} = \inf\{K > 0,\ |f| \leq K\ \mu - \text{a.e.}\}, \quad \text{for } k = \infty.$$

We denote the associated equivalent classes as $\mathbb{L}_k(A, \mu)$ where any two functions coincide for μ-a.e. can not be distinguished. In particular, we write the norm $\|\cdot\|_k$ with $k \in [1, +\infty]$ when $\mu = \lambda$ is the Lebesgue measure. Throughout the paper, C denotes positive numerical constant which may vary from line to line.

3. Estimation based on model selection

Our approach is based on ρ-estimation. For basic ideas that underline the construction of the ρ-estimator, we refer Baraud and Birgé (2018) and Baraud et al. (2017).

3.1. **Main assumption.** Let $\mathcal{M}$ be a finite or countable set. For each $m \in \mathcal{M}$, $\overline{\mathbf{\Gamma}}_m$ stands for a class of measurable functions from $\mathscr{W}$ into I, which we call it *a model*. We begin with an at most countable family $\{\overline{\mathbf{\Gamma}}_m,\ m \in \mathcal{M}\}$ of classes and assume the following.

Assumption 1. *For any $m \in \mathcal{M}$, $\overline{\mathbf{\Gamma}}_m$ is VC-subgraph on $\mathscr{W}$ with dimension not larger than $V_m \geq 1$.*

For completeness, we recall the definition of VC-subgraph. An (open) subgraph of a function $\boldsymbol{\gamma}$ in $\overline{\mathbf{\Gamma}}_m$ is the subset of $\mathscr{W} \times \mathbb{R}$ given by

$$\mathscr{S}_{\boldsymbol{\gamma}} = \{(w,t) \in \mathscr{W} \times \mathbb{R},\ t < \boldsymbol{\gamma}(w)\}.$$

A collection $\overline{\mathbf{\Gamma}}_m$ of measurable functions on $\mathscr{W}$ is VC-subgraph with dimension not larger than V_m if, for any finite subset $\mathcal{S} \subset \mathscr{W} \times \mathbb{R}$ with $|\mathcal{S}| = V_m + 1$, there exists at least one subset S of $\mathcal{S}$ such that for any $\boldsymbol{\gamma} \in \overline{\mathbf{\Gamma}}_m$, S is not the intersection of $\mathcal{S}$ with $\mathscr{S}_{\boldsymbol{\gamma}}$, i.e.

$$S \neq \mathcal{S} \cap \mathscr{S}_{\boldsymbol{\gamma}} \quad \text{whatever } \boldsymbol{\gamma} \in \overline{\mathbf{\Gamma}}_m.$$

In particular, when $\overline{\mathbf{\Gamma}}_m$ is contained in a linear space with finite dimension $d_m \geq 1$ Assumption 1 is fulfilled with $V_m = d_m + 1$. Another property which can be derived from Lemma 2.6.18 of van der Vaart and Wellner (1996) is that if $\overline{\mathbf{\Gamma}}$ is VC-subgraph on a set $\mathscr{W}$ with dimension V and $a, b \in \mathbb{R}$ are fixed numbers, then the classes of functions $\overline{\mathbf{\Gamma}}_a = \left\{\boldsymbol{\gamma} \vee a,\ \boldsymbol{\gamma} \in \overline{\mathbf{\Gamma}}\right\}$ and $\overline{\mathbf{\Gamma}}^b = \left\{\boldsymbol{\gamma} \wedge b,\ \boldsymbol{\gamma} \in \overline{\mathbf{\Gamma}}\right\}$ are also VC-subgraphs on $\mathscr{W}$ with dimension not larger than V. For more properties of the VC-subgraph class of functions, we refer the reader to Section 2.6.2 of van der Vaart and Wellner (1996) and Section 8 of Baraud et al. (2017).

3.2. **Model selection procedure.** We consider $\left\{\overline{\mathbf{\Gamma}}_m,\ m \in \mathcal{M}\right\}$ an at most countable family of models satisfying Assumption 1. To avoid measurability issues, for any $m \in \mathcal{M}$, we take $\mathbf{\Gamma}_m$ a finite or countable subset of $\overline{\mathbf{\Gamma}}_m$ and denote $\mathbf{\Gamma} = \cup_{m \in \mathcal{M}} \mathbf{\Gamma}_m$. Let ψ be the map defined on $[0, +\infty]$ as

$$\psi(x) = \begin{cases} \dfrac{x-1}{x+1} & , \quad x \in [0, +\infty), \\ 1 & , \quad x = +\infty. \end{cases} \tag{5}$$

For any $\boldsymbol{\gamma}, \boldsymbol{\gamma}' \in \mathbf{\Gamma}$, we define the $\mathbf{T}$-statistic as

$$\mathbf{T}(\boldsymbol{X}, \boldsymbol{\gamma}, \boldsymbol{\gamma}') = \sum_{i=1}^{n} \psi\left(\sqrt{\frac{r_{\boldsymbol{\gamma}'(W_i)}(Y_i)}{r_{\boldsymbol{\gamma}(W_i)}(Y_i)}}\right) \tag{6}$$

with the conventions $0/0 = 1$ and $c/0 = +\infty$ for all $c > 0$.

Let Δ be a map from $\mathcal{M}$ to $\mathbb{R}_+$. For each $m \in \mathcal{M}$, we associate it with a nonnegative weight $\Delta(m)$ which satisfies

$$\Sigma = \sum_{m \in \mathcal{M}} e^{-\Delta(m)} < +\infty. \tag{7}$$

In particular, when $\Sigma = 1$, this gives a Bayesian flavour to our procedure by regarding $\Delta(m)$ as a prior distribution on the family $\{\mathbf{\Gamma}_m,\ m \in \mathcal{M}\}$.

Let D_n be a map from $\mathcal{M}$ to $\mathbb{R}_+$ defined as, for any $m \in \mathcal{M}$,

$$D_n(m) = 10^3 V_m \left[9.11 + \log_+\left(\frac{n}{V_m}\right)\right],$$

where V_m stands for the VC dimension of the class $\overline{\mathbf{\Gamma}}_m$. We define the penalty function from $\mathbf{\Gamma}$ to $\mathbb{R}_+$ as

$$\mathbf{pen}(\boldsymbol{\gamma}) = 10^2 \inf_{\{m \in \mathcal{M} | \boldsymbol{\gamma} \in \mathbf{\Gamma}_m\}} \left[D_n(m) + 4.7\Delta(m)\right], \quad \text{for all } \boldsymbol{\gamma} \in \mathbf{\Gamma}. \tag{8}$$

For all $\boldsymbol{\gamma} \in \mathbf{\Gamma}$, we set

$$\boldsymbol{\upsilon}(\boldsymbol{X}, \boldsymbol{\gamma}) = \sup_{\boldsymbol{\gamma}' \in \mathbf{\Gamma}} \left[\mathbf{T}(\boldsymbol{X}, \boldsymbol{\gamma}, \boldsymbol{\gamma}') - \mathbf{pen}(\boldsymbol{\gamma}')\right] + \mathbf{pen}(\boldsymbol{\gamma}). \tag{9}$$

We define $\widehat{\boldsymbol{\gamma}} = \widehat{\boldsymbol{\gamma}}(\boldsymbol{X})$ as any measurable element of the random (and non-void) set

$$\mathcal{E}(\boldsymbol{X}) = \left\{\boldsymbol{\gamma} \in \mathbf{\Gamma} \text{ such that } \boldsymbol{\upsilon}(\boldsymbol{X}, \boldsymbol{\gamma}) \leq \inf_{\boldsymbol{\gamma}' \in \mathbf{\Gamma}} \boldsymbol{\upsilon}(\boldsymbol{X}, \boldsymbol{\gamma}') + 1\right\}. \tag{10}$$

Finally, $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}} = (R_{\widehat{\boldsymbol{\gamma}}}, \ldots, R_{\widehat{\boldsymbol{\gamma}}})$ is our estimator for $\mathbf{R}^\star$.

We comment that the number 1 in (10) does not play any role. Any small number $\delta > 0$ can work for defining our estimator. We hereby choose $\delta = 1$ just for presenting our results in a simple way.

As one can observe from the construction procedure, our estimator depends on the choice of the exponential family in (6), the countable subsets $\mathbf{\Gamma}_m$ of $\overline{\mathbf{\Gamma}}_m$ and the weights $\Delta(m)$ we choose. However, we do not require any information for the distributions of covariates W_i which, therefore, could be unknown. This is one of the feature distinguishing our procedure with the ones Antoniadis et al. (2001), Brown et al. (2010) and Kolaczyk and Nowak (2005) in the literature.

3.3. **The performance of the estimator.**

Theorem 1. *Let $\mathcal{Q}_m = \{\mathbf{R}_{\boldsymbol{\gamma}},\ \boldsymbol{\gamma} \in \mathbf{\Gamma}_m\}$ and $\Xi(m) = D_n(m)/4.7 + \Delta(m)$, for all $m \in \mathcal{M}$. Under Assumption 1, whatever the conditional probabilities $\mathbf{R}^\star = (R_1^\star, \ldots, R_n^\star)$ of Y_i given W_i and the distributions of W_i, the estimator $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}}$ obtained by our model selection procedure in Section 3.2 satisfies for any $\xi > 0$, with a probability at least $1 - \Sigma e^{-\xi}$*

$$\mathbf{h}^2(\mathbf{R}^\star, \mathbf{R}_{\widehat{\boldsymbol{\gamma}}}) \leq \inf_{m \in \mathcal{M}} \left[c_1 \mathbf{h}^2(\mathbf{R}^\star, \mathcal{Q}_m) + c_2\left(\Xi(m) + 1.49 + \xi\right)\right], \tag{11}$$

where $c_1 = 149.8$ and $c_2 = 5013.2$.

The proof of Theorem 1 is postponed to Appendix A. We shall use (11) in the forthcoming sections to solve many model selection problems simultaneously. We need to point out that the result we present here is different with the work in Section 8 of Baraud and Birgé (2018), where they assumed the pairs of random variables $X_i = (W_i, Y_i)$, $i \in \{1, \ldots, n\}$ are i.i.d.. Besides, we study the regression problem in exponential families so that it is more natural to put assumption on the models $\overline{\mathbf{\Gamma}}_m$ of the potential regression function $\boldsymbol{\gamma}^\star$. In this statistical setting, additional work has to be done to understand the performance of the ρ-estimator.

We give some comments on our result here. The numerical constants c_1 and c_2 are independent of the choice of the exponential family. For all $m \in \mathcal{M}$, let us set $\overline{\boldsymbol{\mathcal{Q}}}_m = \{\mathbf{R}_{\boldsymbol{\gamma}},\ \boldsymbol{\gamma} \in \overline{\mathbf{\Gamma}}_m\}$. If for all $m \in \mathcal{M}$, $\boldsymbol{\mathcal{Q}}_m$ is dense in $\overline{\boldsymbol{\mathcal{Q}}}_m$ with respect to the pseudo Hellinger distance $\mathbf{h}$, i.e. $\mathbf{h}(\mathbf{R}^\star, \overline{\boldsymbol{\mathcal{Q}}}_m) = \mathbf{h}(\mathbf{R}^\star, \boldsymbol{\mathcal{Q}}_m)$, (11) is equivalent to

$$\mathbf{h}^2(\mathbf{R}^\star, \mathbf{R}_{\widehat{\boldsymbol{\gamma}}}) \leq \inf_{m \in \mathcal{M}} \left[c_1 \mathbf{h}^2(\mathbf{R}^\star, \overline{\boldsymbol{\mathcal{Q}}}_m) + c_2 \left(\Xi(m) + 1.49 + \xi \right) \right],$$

where we involve the models $\overline{\mathbf{\Gamma}}_m$ into the risk bound of our estimator but not its countable subset $\mathbf{\Gamma}_m$ as we derived in (11). As it was discussed in Section 4.2 of Baraud and Birgé (2018), this is exact the case when $\mathbf{\Gamma}_m$ is a dense subset of $\overline{\mathbf{\Gamma}}_m$ for the topology of pointwise convergence for all $m \in \mathcal{M}$.

An integration of (11) with respect to ξ leads to

$$\mathbb{E}\left[\mathbf{h}^2(\mathbf{R}^\star, \mathbf{R}_{\widehat{\boldsymbol{\gamma}}}) \right] \leq \inf_{m \in \mathcal{M}} \left[c_1 \mathbf{h}^2(\mathbf{R}^\star, \boldsymbol{\mathcal{Q}}_m) + c_2 \left(\Xi(m) + \Sigma + 1.49 \right) \right]. \tag{12}$$

We note from (12) that the risk of the estimator $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}}$ is bounded, up to a constant depending on Σ, by the infimum over the whole family $\mathcal{M}$ of the quantity summing up the distance from each $\boldsymbol{\mathcal{Q}}_m$ to $\mathbf{R}^\star$, the complexity of each $\overline{\mathbf{\Gamma}}_m$ (up to a logarithmic factor) and the associated weight $\Delta(m)$. The magnitude of the bias term and the complexity term is of the optimal order so that if for all $m \in \mathcal{M}$, the weight function $\Delta(m)$ is chosen to be not larger than V_m (up to a logarithmic factor), we are able to select the model achieving the best trade-off between approximation and model's complexity among the collection $\mathcal{M}$.

Moreover, the bias term $\mathbf{h}(\mathbf{R}^\star, \boldsymbol{\mathcal{Q}}_m)$ in (12) accounts for the robustness property of our estimator with respect to the possible model misspecification and data contamination. To illustrate it simply, let us focus on each single $\mathbf{\Gamma}_m$ and assume the weight $\Delta(m)$ has been assigned such that $\Delta(m) \lesssim D_n(m)$. If $\mathbf{\Gamma}_m$ is exact, i.e. $\mathbf{R}^\star = \mathbf{R}_{\boldsymbol{\gamma}^\star}$ with $\boldsymbol{\gamma}^\star \in \mathbf{\Gamma}_m$, up to a constant, the risk of the estimator $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}}$ will be smaller than $V_m \left[1 + \log_+ (n/V_m) \right]$. If it is not the case, the risk involves an additional bias term $\mathbf{h}^2(\mathbf{R}^\star, \boldsymbol{\mathcal{Q}}_m)$ due to a potential model misspecification or data contamination. However, as long as this bias term remains small compared to $V_m \left[1 + \log_+ (n/V_m) \right]$, the performance of our estimator will not deteriorate much as the case when $\mathbf{\Gamma}_m$ is exact.

In the situation where the covariates W_i are truly i.i.d. with a common distribution P_W and $R_i^\star = R^\star$ for all $i \in \{1, \dots, n\}$, we deduce from (12) that for any $R^\star$ and P_W, our estimator $R_{\widehat{\gamma}}$ satisfies

(13)
$$\mathbb{E}\left[h^2(R^\star, R_{\widehat{\gamma}})\right] \leq c_2\left(c_3 + \Sigma\right) \inf_{m \in \mathcal{M}} \left[h^2(R^\star, \mathscr{Q}_m) + \frac{\Delta(m)}{n} + \frac{V_m}{n} L_n(m)\right],$$

where $c_3 = 1939.8$, $\mathscr{Q}_m = \{R_{\boldsymbol{\gamma}},\ \boldsymbol{\gamma} \in \boldsymbol{\Gamma}_m\}$ and $L_n(m) = 1 + \log_+\left(n/V_m\right)$.

4. Adaptation to anisotropic Besov spaces

In this section, we assume the covariates W_i are truly i.i.d. on $\mathscr{W} = [0,1]^d$, $d \geq 1$ with a common distribution P_W and $R_i^\star = R^\star$ for all $i \in \{1, \dots, n\}$ and consider adaptive estimation in exponential families. The problem is stated as follows.

Let $0 < p, q \leq \infty$, $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_d) \in (\mathbb{R}_+^*)^d$ and $R \in \mathbb{R}_+^*$. We denote $B_{p,q}^{\boldsymbol{\alpha}}([0,1]^d, R)$ as the anisotropic Besov ball which gathers all the functions f in the anisotropic Besov space $B_{p,q}^{\boldsymbol{\alpha}}([0,1]^d)$ with (quasi-) semi-norm $|f|_{\boldsymbol{\alpha},p,q} < R$. Including Hölder and Sobolev spaces, Besov space is a considerable general function space. It can also capture the spatial inhomogeneity of the smoothness property as discussed by Suzuki and Nitanda (2021). For readers who concern the definitions, we refer Chapter 5 of Triebel (2006) and Hochmuch (2002) which gives a detailed introduction restricted to $d = 2$ but can be generalized easily. Similarly to the isotropic case, the d-dimensional parameter $\boldsymbol{\alpha}$ indicates the smooth property in each direction $j \in \{1, \dots, d\}$. More precisely, for all functions $f \in B_{p,q}^{\boldsymbol{\alpha}}([0,1]^d)$, if α_j is large, then f is smooth to the j-th direction.

For a given interval $[v_-, v_+] \subset I$ with $v_- < v_+$, the notation $B_{p,q}^{\boldsymbol{\alpha}}(R, v_-, v_+)$ stands for the collection of functions $f \in B_{p,q}^{\boldsymbol{\alpha}}([0,1]^d, R)$ with $f(\boldsymbol{w}) \in [v_-, v_+]$ for all $\boldsymbol{w} \in [0,1]^d$. We assume that the regression function $\boldsymbol{\gamma}^\star \in B_{p,q}^{\boldsymbol{\alpha}}(R, v_-, v_+)$. Our aim, in this section, is to design a specific procedure for estimating this $\boldsymbol{\gamma}^\star$ without assuming the parameters $\boldsymbol{\alpha}$, p and R to be known.

4.1. Models construction.

We begin with introducing the conception of hyperrectangle. Given $s_j \in \mathbb{N}$, $1 \leq j \leq d$, for any $k_j \in \Psi(s_j) = \{0, \dots, 2^{s_j} - 1\}$, we set

$$I_j(k_j) = \begin{cases} \left[0, 2^{-s_j}\right] & , \quad k_j = 0, \\ \left(k_j 2^{-s_j}, (k_j+1)2^{-s_j}\right] & , \quad k_j = 1, \dots, 2^{s_j} - 1. \end{cases} \tag{14}$$

We call a hyperrectangle by any subset of $[0,1]^d$ of the form $\prod_{j=1}^d I_j(k_j)$. Given a vector $\boldsymbol{s} = (s_1, \dots, s_d) \in \mathbb{N}^d$, we denote $M_{\boldsymbol{s}}^{\mathcal{B},d}$ the resulted partition of $[0,1]^d$ into the union of hyperrectangles $\cup_{(k_1,\dots,k_d) \in \Psi(s_1) \times \cdots \times \Psi(s_d)} \prod_{j=1}^d I_j(k_j)$.

We take $\mathcal{M} = \mathbb{N}^d \times \mathbb{N}$. Given $(\boldsymbol{s}, r) \in \mathcal{M}$, we define $\overline{\boldsymbol{S}}^{\mathcal{B},d}_{(\boldsymbol{s},r)}$ as the space of piecewise polynomial functions on $[0,1]^d$, where on each hyperrectangle $\prod_{j=1}^d I_j(k_j)$, $\boldsymbol{\gamma} \in \overline{\boldsymbol{S}}^{\mathcal{B},d}_{(\boldsymbol{s},r)}$ is a polynomial in d variables of degree at most r for each variable. This is to say given $(\boldsymbol{s}, r) \in \mathcal{M}$, for any $(\overline{k}_1, \ldots, \overline{k}_d) \in \Psi(s_1) \times \cdots \times \Psi(s_d)$, any $\boldsymbol{\gamma} \in \overline{\boldsymbol{S}}^{\mathcal{B},d}_{(\boldsymbol{s},r)}$ is of the form for all $\boldsymbol{w} = (w_1, \ldots, w_d) \in \prod_{j=1}^d I_j(\overline{k}_j)$

$$\boldsymbol{\gamma}(\boldsymbol{w}) = \sum_{(r_1,\ldots,r_d)\in\{0,\ldots,r\}^d} \gamma_{(r_1,\ldots,r_d)} \prod_{j=1}^d {w_j}^{r_j}, \tag{15}$$

where $\gamma_{(r_1,\ldots,r_d)} \in \mathbb{R}$, for all $0 \leq r_j \leq r$, $1 \leq j \leq d$.

Recall that in our setting $\boldsymbol{\gamma}^\star$ takes values in some non-trivial interval I which may vary from the choice of the exponential family and the choice of parametrization. A few examples are given in Section 2.1 of Baraud and Chen (2020) when the exponential family is parametrized in its natural form. To estimate $\boldsymbol{\gamma}^\star$, we assume that we have a prior information of $v_-, v_+ \in \mathbb{R}$ such that the regression function $\boldsymbol{\gamma}^\star$ with values in $[v_-, v_+] \subset I$. For each $(\boldsymbol{s}, r) \in \mathcal{M}$, we define $\overline{\boldsymbol{\Gamma}}^{\mathcal{B},d}_{(\boldsymbol{s},r)} = \left\{(\boldsymbol{\gamma} \vee v_-) \wedge v_+,\ \boldsymbol{\gamma} \in \overline{\boldsymbol{S}}^{\mathcal{B},d}_{(\boldsymbol{s},r)}\right\}$ and the family of models is given by $\left\{\overline{\boldsymbol{\Gamma}}^{\mathcal{B},d}_{(\boldsymbol{s},r)},\ (\boldsymbol{s}, r) \in \mathcal{M}\right\}$. For each $\overline{\boldsymbol{S}}^{\mathcal{B},d}_{(\boldsymbol{s},r)}$, we take its countable subset $\boldsymbol{S}^{\mathcal{B},d}_{(\boldsymbol{s},r)}$ as the collection of functions of the same form in (15) apart from restricting $\gamma_{(r_1,\ldots,r_d)} \in \mathbb{Q}$, for all $0 \leq r_j \leq r$, $1 \leq j \leq d$ and define $\boldsymbol{\Gamma}^{\mathcal{B},d}_{(\boldsymbol{s},r)} = \left\{(\boldsymbol{\gamma} \vee v_-) \wedge v_+,\ \boldsymbol{\gamma} \in \boldsymbol{S}^{\mathcal{B},d}_{(\boldsymbol{s},r)}\right\}$.

Lemma 1. *For any $d \in \mathbb{N}^*$, $r \in \mathbb{N}$ and $\boldsymbol{s} \in \mathbb{N}^d$, $\boldsymbol{S}^{\mathcal{B},d}_{(\boldsymbol{s},r)}$ is dense in $\overline{\boldsymbol{S}}^{\mathcal{B},d}_{(\boldsymbol{s},r)}$ and $\boldsymbol{\Gamma}^{\mathcal{B},d}_{(\boldsymbol{s},r)}$ is dense in $\overline{\boldsymbol{\Gamma}}^{\mathcal{B},d}_{(\boldsymbol{s},r)}$ with respect to the supremum norm $\|\cdot\|_\infty$.*

For any $(\boldsymbol{s}, r) \in \mathcal{M}$, since $M^{\mathcal{B},d}_{\boldsymbol{s}}$ is a partition of $[0,1]^d$ with $\prod_{j=1}^d 2^{s_j}$ hyperrectangles and on each hyperrectangle the space of functions is spanned by $(r+1)^d$ basis, $\overline{\boldsymbol{S}}^{\mathcal{B},d}_{(\boldsymbol{s},r)}$ is a $(r+1)^d \prod_{j=1}^d 2^{s_j}$ dimensional vector space. By the properties of VC-subgraph we introduced in Section 3, for any $(\boldsymbol{s}, r) \in \mathcal{M}$, $\overline{\boldsymbol{\Gamma}}^{\mathcal{B},d}_{(\boldsymbol{s},r)}$ is a VC-subgraph on $\mathscr{W}$ with dimension not lager than $(r+1)^d \prod_{j=1}^d 2^{s_j} + 1$ which fulfills Assumption 1 with

$$V_{(\boldsymbol{s},r)} = (r+1)^d \prod_{j=1}^d 2^{s_j} + 1. \tag{16}$$

For each $(\boldsymbol{s}, r) \in \mathcal{M}$, we associate it with the weight

$$\Delta(\boldsymbol{s}, r) = \log(8d) \prod_{j=1}^d 2^{s_j} + r. \tag{17}$$

We have the following result which shows inequality (7) is satisfied with the weights defined by (17).

Lemma 2. *For each $(\boldsymbol{s}, r) \in \mathcal{M}$, let the weight be assigned by (17). Then*

$$\sum_{(\boldsymbol{s},r)\in\mathcal{M}} e^{-\Delta(\boldsymbol{s},r)} \leq \frac{e}{e-1}.$$

We denote $M^{\mathcal{B},d} = \cup_{\boldsymbol{s}\in\mathbb{N}^d} M_{\boldsymbol{s}}^{\mathcal{B},d}$. Given a partition $\pi \in M^{\mathcal{B},d}$ without knowing the specific values of $(s_1, \ldots, s_d)$, sometimes it is useful to introduce an alternative notation $\overline{\boldsymbol{\Gamma}}_{(\pi,r)}^{\mathcal{B},d} = \left\{(\boldsymbol{\gamma} \vee v_-) \wedge v_+,\ \boldsymbol{\gamma} \in \overline{\boldsymbol{S}}_{(\pi,r)}^{\mathcal{B},d}\right\}$ for $\overline{\boldsymbol{\Gamma}}_{(\boldsymbol{s},r)}^{\mathcal{B},d}$, where $\overline{\boldsymbol{S}}_{(\pi,r)}^{\mathcal{B},d}$ characterises the space of piecewise polynomial functions on $[0,1]^d$ such that on each hyperrectangle of π, any $\boldsymbol{\gamma} \in \overline{\boldsymbol{S}}_{(\pi,r)}^{\mathcal{B},d}$ is a polynomial in d variables of degree not larger than r for each variable. Similarly, the VC dimension bound for the class of functions $\overline{\boldsymbol{\Gamma}}_{(\pi,r)}^{\mathcal{B},d}$ on $\mathscr{W}$ is given by

$$V_{(\pi,r)} = (r+1)^d|\pi| + 1, \tag{18}$$

where $|\pi|$ denotes the cardinality of hyperrectangles given by the partition π of $[0,1]^d$. Under this new notation, the weight associated to each $(\pi, r) \in M^{\mathcal{B},d} \times \mathbb{N}$ can be deduced from (17) as

$$\Delta(\pi, r) = \log(8d)|\pi| + r. \tag{19}$$

4.2. **Adaptivity result.** Before deriving the risk bound for our estimator based on the constructed family $\left\{\boldsymbol{\Gamma}_{(\boldsymbol{s},r)}^{\mathcal{B},d},\ (\boldsymbol{s}, r) \in \mathcal{M}\right\}$, we first discuss the parametrization issue of the exponential family. As it has been explained in Section 4.1 and 4.2 of Baraud and Chen (2020), the parametrization of the exponential family influences the converge rate of $\widehat{\boldsymbol{\gamma}}$ to $\boldsymbol{\gamma}^\star$. For example, when $d = 1$ one can see from Section 4.1 of Baraud and Chen (2020) that if we parametrize exponential families by their mean, Poisson regression achieves much slower rate than the Gaussian case under the same α-Hölder smoothness assumption on $\boldsymbol{\gamma}^\star$ with $\alpha \in (0, 1]$. However, there do exist ways of parametrization such that the same rate of convergence can be achieved uniformly regardless the choice of the exponential family. We assume the following holds.

Assumption 2. *The exponential family $\widetilde{\mathscr{Q}} = \{R_\gamma,\ \gamma \in I\}$ has been parametrized in the way that there exists a constant $\kappa > 0$ such that*

$$h(R_\gamma, R_{\gamma'}) \leq \kappa|\gamma - \gamma'| \quad \textit{for all} \quad \gamma, \gamma' \in I.$$

Let us remark that, by Proposition 2 of Baraud and Chen (2020), Assumption 2 is fulfilled with $\kappa = 1$ when the exponential family is parametrized by $\gamma = \upsilon(\theta)$, where θ is the natural parameter and υ satisfies $\upsilon'(\theta) = \sqrt{A''(\theta)/8}$ with the function A defined as

$$A(\theta) = \log\left[\int_{\mathscr{Y}} e^{\theta T(y)} d\nu(y)\right].$$

For any $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) \in (\mathbb{R}_+^*)^d$, we denote $\alpha_{\min} = \min_{1 \leq j \leq d} \alpha_j$ and $\overline{\alpha}$ the harmonic mean of $\alpha_1, \ldots, \alpha_d$, i.e.

$$\overline{\alpha} = \left(\frac{1}{d} \sum_{j=1}^{d} \frac{1}{\alpha_j} \right)^{-1}.$$

With the family of models $\left\{ \overline{\mathbf{\Gamma}}^{\mathcal{B},d}_{(\boldsymbol{s},r)},\ (\boldsymbol{s},r) \in \mathcal{M} \right\}$ defined in Section 4.1, the associated countable subsets $\left\{ \mathbf{\Gamma}^{\mathcal{B},d}_{(\boldsymbol{s},r)},\ (\boldsymbol{s},r) \in \mathcal{M} \right\}$ and the weights defined by (17), we are now able to apply the model selection procedure introduced in Section 3.2 to estimate $\boldsymbol{\gamma}^\star$. The following result shows that under Assumption 2, the resulted estimator $\widehat{\boldsymbol{\gamma}}(\boldsymbol{X})$ based on $\left\{ \mathbf{\Gamma}^{\mathcal{B},d}_{(\boldsymbol{s},r)},\ (\boldsymbol{s},r) \in \mathcal{M} \right\}$ is adapted to the possible anisotropy over a wide range of the anisotropic Besov spaces with a risk bound of order $n^{-2\overline{\alpha}/(2\overline{\alpha}+d)}$ up to a logarithmic factor with respect to the distance $d(\boldsymbol{\gamma}^\star, \widehat{\boldsymbol{\gamma}}) = h^2\left(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}}\right)$. One nice feature is that this risk bound is independent of the choice of the exponential family.

Corollary 1. *Under Assumption 2, whatever the distribution of W, the estimator $\widehat{\boldsymbol{\gamma}}(\boldsymbol{X})$ given by the model selection procedure in Section 3.2 over the family $\left\{ \mathbf{\Gamma}^{\mathcal{B},d}_{(\boldsymbol{s},r)},\ (\boldsymbol{s},r) \in \mathbb{N}^d \times \mathbb{N} \right\}$ with the weights defined by (17) satisfies for all $R > 0$, $p > 0$ and $\boldsymbol{\alpha} \in (\mathbb{R}_+^*)^d$ such that $\overline{\alpha}/d > 1/p$,*

$$\sup_{\boldsymbol{\gamma}^\star \in B^{\boldsymbol{\alpha}}_{p,q}(R, v_-, v_+)} \mathbb{E}\left[h^2\left(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}}\right) \right] \leq C_{\kappa,d,\boldsymbol{\alpha},p} \left(R^{\frac{2d}{d+2\overline{\alpha}}} n^{-\frac{2\overline{\alpha}}{d+2\overline{\alpha}}} + \frac{1}{n} \right) (1 + \log n),$$

where $q = \infty$ if $0 < p \leq 1$ or $p \geq 2$ and $q = p$ if $1 < p < 2$, $C_{\kappa,d,\boldsymbol{\alpha},p}$ is a constant depending on $\kappa, d, \boldsymbol{\alpha}, p$ only.

The proof of Corollary 1 is postponed to Appendix A. We hereby give some comments on this result. First, Corollary 1 in fact holds for any $0 < q \leq \infty$, if $0 < p \leq 1$ or $p \geq 2$ and $0 < q \leq p$, if $1 < p < 2$ as a consequence of embedding the anisotropic Besov spaces to some bigger spaces. We will not discuss too much on this direction but refer the reader to Section 2.3 of Akakpo (2012). Second, as it has been discussed by Section 2.1 of Suzuki and Nitanda (2021), the parameter p plays a role of controlling the spatial inhomogeneity of the smoothness. In particular, when $p = \infty$, the smoothness is ensured uniformly. Our result, therefore, is also adapted to $\boldsymbol{\gamma}^\star$ with potentially inhomogeneous smoothness. Third, the rate is optimal up to a logarithmic factor in the minimax sense at least when $d = 1$ as it has been shown in Proposition 4 of Baraud and Chen (2020). Finally, the condition $\overline{\alpha}/d > 1/p$ appearing in the result is more strict than the usual one which only requires $\overline{\alpha}/d > (1/p - 1/2)_+$. This is because we do not make any assumption on the distribution of the covariate W. Therefore, we bound the approximation bias with respect to the sup-norm $\|\cdot\|_\infty$. As one can see from the proof of Corollary 1, this bias bound can be reconsidered if the specific distribution of the covariate W is given. In the particular case when the probability measure P_W admits a density $P_W = p_W \cdot \lambda$ with

respect to the Lebesgue measure λ and $\|p_W\|_\infty \leq K$ (i.e. the probability measure P_W is equivalent to the Lebesgue probability on $\mathscr{W} = [0,1]^d$), we only need to require the usual condition $\overline{\alpha}/d > (1/p - 1/2)_+$ to obtain the same rate in Corollary 1, where the numerical constant depends on K, κ, d, $\boldsymbol{\alpha}$ and p.

5. Model selection under structural assumptions

In the last section, we have seen that when the covariates W_i are truly i.i.d. on $[0,1]^d$ and $R_i^\star = R_{\boldsymbol{\gamma}^\star}$ for all $i \in \{1,\dots,n\}$ with $\boldsymbol{\gamma}^\star \in B_{p,q}^{\boldsymbol{\alpha}}(R, v_-, v_+)$, the estimator $\widehat{\boldsymbol{\gamma}}(\boldsymbol{X})$ obtained from our model selection procedure based on $\left\{\boldsymbol{\Gamma}_{(\boldsymbol{s},r)}^{\mathcal{B},d},\ (\boldsymbol{s},r) \in \mathcal{M}\right\}$ achieves the converge rate $n^{-2\overline{\alpha}/(d+2\overline{\alpha})}$ adaptively. When the value of d is large, this rate becomes slow, which is, as a well-known phenomenon, called the curse of dimensionality. To circumvent it, we impose structural assumptions on $\boldsymbol{\gamma}^\star$ in this section and consider additional models to implement our procedure. We mainly discuss two examples of the structural assumptions: generalized additive structure and multiple index structure.

We begin with setting some notations. Let $k \in \mathbb{N}^*$ and $\boldsymbol{w} = (w_1,\dots,w_k) \in [0,1]^k$. For a vector $\boldsymbol{\alpha} = (\alpha_1,\dots,\alpha_k) \in (\mathbb{R}_+^*)^k$ with $\alpha_j = r_j + \alpha_j'$, $r_j \in \mathbb{N}$ and $\alpha_j' \in (0,1]$ for $j \in \{1,\dots,k\}$, Hölder space $\mathcal{H}^{\boldsymbol{\alpha}}([0,1]^k)$ denotes the collection of functions f on $[0,1]^k$ satisfying for any $(w_1,\dots,w_{j-1},w_{j+1},\dots,w_k) \in [0,1]^{k-1}$ and all $x, y \in [0,1]$

$$\left|\partial_j^{r_j} f(w_1,\dots,x,\dots,w_k) - \partial_j^{r_j} f(w_1,\dots,y,\dots,w_k)\right| \leq L(f)|x-y|^{\alpha_j'},$$

where $\partial_j^{r_j} f$ denotes the r_j-th order partial derivative of the function f on the j-th component. We define the anisotropic Hölder class $\mathcal{H}^{\boldsymbol{\alpha}}([0,1]^k, L)$ as the collection of all the functions $f \in \mathcal{H}^{\boldsymbol{\alpha}}([0,1]^k)$ with $L(f) + \inf \overline{L} \leq L$, where the infimum runs among all $\overline{L}$ such that

$$|f(\boldsymbol{w}) - f(\boldsymbol{w}')| \leq \overline{L} \sum_{j=1}^k |w_j - w_j'|^{\alpha_j \wedge 1}, \text{ for all } \boldsymbol{w}, \boldsymbol{w}' \in [0,1]^k$$

and define $\mathcal{H}^{\boldsymbol{\alpha}}(L, v_-, v_+)$ as the collection of all functions $f \in \mathcal{H}^{\boldsymbol{\alpha}}([0,1]^k, L)$ taking values in $[v_-, v_+] \subset I$ with $v_- < v_+$.

Given $t_j \in \mathbb{N}^*$, $1 \leq j \leq k$, for any $h_j \in \Phi(t_j) = \{0,\dots,t_j - 1\}$, we define

$$I_j'(h_j) = \begin{cases} [0, 1/t_j] & , \quad h_j = 0, \\ (h_j/t_j, (h_j+1)/t_j] & , \quad h_j = 1,\dots,t_j - 1. \end{cases} \tag{20}$$

For a given $k \in \mathbb{N}^*$ and $\boldsymbol{t} = (t_1, \ldots, t_k) \in (\mathbb{N}^*)^k$, we denote $M_{\boldsymbol{t}}^{\mathcal{H},k}$ the resulted partition of $[0,1]^k$ into the union of $\prod_{j=1}^k t_j$ hyperrectangles

$$\cup_{(h_1,\ldots,h_k)\in\Phi(t_1)\times\cdots\times\Phi(t_k)} \prod_{j=1}^k I_j'(h_j),$$

where on j-th direction the interval $[0,1]$ is divided into t_j regular subintervals, for $j \in \{1,\ldots,k\}$. For any $k \in \mathbb{N}^*$, $\boldsymbol{t} \in (\mathbb{N}^*)^k$ and $r \in \mathbb{N}$, we denote $\overline{\boldsymbol{S}}_{(\boldsymbol{t},r)}^{\mathcal{H},k}$ the space of piecewise polynomial functions f on $[0,1]^k$ such that the restriction of f to each hyperrectangle is a polynomial in k variables of degree not larger than r for each variable and $\boldsymbol{S}_{(\boldsymbol{t},r)}^{\mathcal{H},k}$ the collection of functions with the same form as the ones belonging to $\overline{\boldsymbol{S}}_{(\boldsymbol{t},r)}^{\mathcal{H},k}$ apart from restricting the coefficients in front of the polynomial basis to be rational numbers. With a similar argument as the proof of Lemma 1, the following result is easy to obtain.

Lemma 3. *For any $k \in \mathbb{N}^*$, $\boldsymbol{t} \in (\mathbb{N}^*)^k$ and $r \in \mathbb{N}$, $\boldsymbol{S}_{(\boldsymbol{t},r)}^{\mathcal{H},k}$ is dense in $\overline{\boldsymbol{S}}_{(\boldsymbol{t},r)}^{\mathcal{H},k}$ with respect to the supremum norm $\|\cdot\|_\infty$.*

5.1. **Generalized additive structure.** Generalized additive functions, as a classical structural assumption, have been considered in many statistical literatures. Let $\alpha, L \in \mathbb{R}_+^*$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d) \in (\mathbb{R}_+^*)^d$, $\mathbf{p} = (p_1, \ldots, p_d) \in (\mathbb{R}_+^*)^d$ and $\mathbf{R} = (R_1, \ldots, R_d) \in (\mathbb{R}_+^*)^d$. We denote $\mathcal{F}_{[v_-,v_+]}(\alpha, \boldsymbol{\beta}, \mathbf{p}, L, \mathbf{R})$ the collection of functions $\boldsymbol{\gamma} : [0,1]^d \to [v_-, v_+] \subset I$ of the following form

$$\boldsymbol{\gamma}(\boldsymbol{w}) = f\left(\sum_{j=1}^d g_j(w_j)\right), \quad \text{for all } \boldsymbol{w} = (w_1, \ldots, w_d) \in [0,1]^d,$$

where $f \in \mathcal{H}^\alpha(L, v_-, v_+)$ and $g_j \in B_{p_j,p_j}^{\beta_j}([0,1], R_j)$ taking values in $[0, 1/d]$, for $j \in \{1, \ldots, d\}$.

We assume the regression function $\boldsymbol{\gamma}^\star \in \mathcal{F}_{[v_-,v_+]}(\alpha, \boldsymbol{\beta}, \mathbf{p}, L, \mathbf{R})$ but without the knowledge of α, $\boldsymbol{\beta}$, $\mathbf{p}$, L and $\mathbf{R}$. To estimate $\boldsymbol{\gamma}^\star$ by our model selection procedure, we need to first build suitable approximation models for the class of functions $\mathcal{F}_{[v_-,v_+]}(\alpha, \boldsymbol{\beta}, \mathbf{p}, L, \mathbf{R})$.

To approximate the Besov class of functions $B_{p_j,p_j}^{\beta_j}([0,1], R_j)$, we consider the family $\left\{\overline{\boldsymbol{S}}_{(s,r)}^{\mathcal{B},1},\ (s,r) \in \mathbb{N} \times \mathbb{N}\right\}$ introduced in Section 4.1 taking $d = 1$. We recall that the functions belonging to the above family are built based on the collection of particular partitions $M^{\mathcal{B},1} = \cup_{s\in\mathbb{N}} M_s^{\mathcal{B},1}$. Therefore, we can rewrite the family in an alternative way $\left\{\overline{\boldsymbol{S}}_{(\pi,r)}^{\mathcal{B},1},\ (\pi,r) \in M^{\mathcal{B},1} \times \mathbb{N}\right\}$. To approximate the Hölder class of functions $\mathcal{H}^\alpha([0,1], L)$ with values in $[v_-, v_+]$, we consider the family $\left\{\overline{\boldsymbol{\Gamma}}_{(t,r)}^{\mathcal{H},1},\ (t,r) \in \mathbb{N}^* \times \mathbb{N}\right\}$, where $\overline{\boldsymbol{\Gamma}}_{(t,r)}^{\mathcal{H},1} = \left\{(\boldsymbol{\gamma} \vee v_-) \wedge v_+,\ \boldsymbol{\gamma} \in \overline{\boldsymbol{S}}_{(t,r)}^{\mathcal{H},1}\right\}$.

For any $r \in \mathbb{N}$, $t \in \mathbb{N}^*$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_d) \in (M^{\mathcal{B},1})^d$, we define $\overline{\boldsymbol{\Gamma}}^A_{(\boldsymbol{\pi},t,r)}$ the collection of all the functions $\boldsymbol{\gamma}$ on $\mathscr{W} = [0,1]^d$ of the form

$$\boldsymbol{\gamma}(\boldsymbol{w}) = f\left[(g(\boldsymbol{w}) \vee 0) \wedge 1\right], \quad \text{for all } \boldsymbol{w} = (w_1, \ldots, w_d) \in [0,1]^d, \tag{21}$$

where $g(\boldsymbol{w}) = \sum_{j=1}^d g_j(w_j)$ with $g_j \in \overline{\boldsymbol{S}}^{\mathcal{B},1}_{(\pi_j,r)}$, for $j \in \{1, \ldots, d\}$ and $f \in \overline{\boldsymbol{\Gamma}}^{\mathcal{H},1}_{(t,r)}$. The following result reveals the upper bound of the VC dimension for the class of functions $\overline{\boldsymbol{\Gamma}}^A_{(\boldsymbol{\pi},t,r)}$.

Proposition 1. *Given $r \in \mathbb{N}$, $t \in \mathbb{N}^*$ and $\boldsymbol{\pi} \in (M^{\mathcal{B},1})^d$, the class of functions $\overline{\boldsymbol{\Gamma}}^A_{(\boldsymbol{\pi},t,r)}$ is a VC-subgraph on $[0,1]^d$ with dimension*

$$V^A_{(\boldsymbol{\pi},t,r)} \leq 2 + \left[t(r+1) + 2\sum_{j=1}^d |\pi_j|(r+1)\right] \log_2\left[4eU \log_2(2eU)\right],$$

where $U = t + r + 2$.

The proof is postponed to Appendix C. For each $r \in \mathbb{N}$, $t \in \mathbb{N}^*$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_d) \in (M^{\mathcal{B},1})^d$, we take the countable subset $\boldsymbol{\Gamma}^A_{(\boldsymbol{\pi},t,r)}$ of $\overline{\boldsymbol{\Gamma}}^A_{(\boldsymbol{\pi},t,r)}$ defined as

$$\boldsymbol{\Gamma}^A_{(\boldsymbol{\pi},t,r)} = \left\{f\left[(g \vee 0) \wedge 1\right],\ f \in \boldsymbol{\Gamma}^{\mathcal{H},1}_{(t,r)},\ g_j \in \boldsymbol{S}^{\mathcal{B},1}_{(\pi_j,r)},\ j = 1, \ldots, d\right\},$$

where $\boldsymbol{S}^{\mathcal{B},1}_{(\pi_j,r)}$ is the rational version of $\overline{\boldsymbol{S}}^{\mathcal{B},1}_{(\pi_j,r)}$ which has been introduced in Section 4.1, $\boldsymbol{\Gamma}^{\mathcal{H},1}_{(t,r)} = \left\{(\boldsymbol{\gamma} \vee v_-) \wedge v_+,\ \boldsymbol{S}^{\mathcal{H},1}_{(t,r)}\right\}$ and $g(\boldsymbol{w}) = \sum_{j=1}^d g_j(w_j)$, for all $\boldsymbol{w} = (w_1, \ldots, w_d) \in [0,1]^d$.

Let $\mathcal{M} = (M^{\mathcal{B},1})^d \times \mathbb{N}^* \times \mathbb{N}$. For any $(\boldsymbol{\pi}, t, r) \in (M^{\mathcal{B},1})^d \times \mathbb{N}^* \times \mathbb{N}$, we associate it with the weight

$$\Delta(\boldsymbol{\pi}, t, r) = 3\log 2\left(\sum_{j=1}^d |\pi_j|\right) + r + t. \tag{22}$$

The following result shows inequality (7) is satisfied with the weights defined by (22).

Lemma 4. *With the weights defined by (22), we have*

$$\sum_{(\boldsymbol{\pi},t,r) \in (M^{\mathcal{B},1})^d \times \mathbb{N}^* \times \mathbb{N}} e^{-\Delta(\boldsymbol{\pi},t,r)} \leq \frac{e}{e-1}.$$

With Proposition 1 and Lemma 4, we can apply the model selection procedure introduced in Section 3.2 and obtain the following.

Corollary 2. *Under Assumption 2, no matter what the distribution of W is, the estimator $\widehat{\boldsymbol{\gamma}}(\boldsymbol{X})$ given by the model selection procedure in Section 3.2*

over $\left\{\mathbf{\Gamma}^A_{(\boldsymbol{\pi},t,r)},\ (\boldsymbol{\pi},t,r)\in (M^{\mathcal{B},1})^d\times\mathbb{N}^*\times\mathbb{N}\right\}$ *with the weights defined by (22) satisfies for all* $\alpha, L\in\mathbb{R}^*_+$ *and* $\boldsymbol{\beta},\mathbf{p},\mathbf{R}\in(\mathbb{R}^*_+)^d$ *such that* $\beta_j>1/p_j$

$$\sup_{\boldsymbol{\gamma}^\star\in\mathcal{F}_{[v_-,v_+]}(\alpha,\boldsymbol{\beta},\mathbf{p},L,\mathbf{R})} C'_{\kappa,d,\alpha,\boldsymbol{\beta},\mathbf{p}}\mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star},R_{\widehat{\boldsymbol{\gamma}}})\right]$$

$$\leq\left\{\left[\sum_{j=1}^d\left(LR_j^{\alpha\wedge1}\right)^{\frac{2}{2(\alpha\wedge1)\beta_j+1}}n^{-\frac{2(\alpha\wedge1)\beta_j}{2(\alpha\wedge1)\beta_j+1}}\right]+L^{\frac{2}{2\alpha+1}}n^{-\frac{2\alpha}{2\alpha+1}}+\frac{1}{n}\right\}\mathcal{L}_n^2, \tag{23}$$

where $\mathcal{L}_n=\log n\vee\log L^2\vee1$ *and* $C'_{\kappa,d,\alpha,\boldsymbol{\beta},\mathbf{p}}$ *is a constant depending on* κ, d, α, $\boldsymbol{\beta}$ *and* $\mathbf{p}$.

Corollary 2 tells that in the ideal situation $\boldsymbol{\gamma}^\star\in\mathcal{F}_{[v_-,v_+]}(\alpha,\boldsymbol{\beta},\mathbf{p},L,\mathbf{R})$ for some α, $\boldsymbol{\beta}$, $\mathbf{p}$, L and $\mathbf{R}$, the converge rate of the estimator is independent of d which entails the procedure does not suffer from the curse of dimensionality. When $R^\star\neq R_{\boldsymbol{\gamma}^\star}$ or $\boldsymbol{\gamma}^\star$ exists but does not belong to any $\mathcal{F}_{[v_-,v_+]}(\alpha,\boldsymbol{\beta},\mathbf{p},L,\mathbf{R})$, a bias term will be added into the risk bound in Corollary 2. However, as long as the bias term is not too large compared to the quantity on the right hand side of (23), the accuracy of the resulted estimator $\widehat{\boldsymbol{\gamma}}(\boldsymbol{X})$ remains the same magnitude as the ideal case which confirms the robustness of our estimator.

5.2. **Multiple index structure.** Let $\mathcal{C}_d$ be the unit ball for the ℓ_1-norm, i.e.

$$\mathcal{C}_d=\left\{(c_1,\ldots,c_d)\in\mathbb{R}^d,\ \sum_{j=1}^d|c_j|\leq1\right\}.$$

For some known $l\in\mathbb{N}^*$ (typically $l\leq d$), we denote $\mathcal{G}_{[v_-,v_+]}(\boldsymbol{\alpha},L)$ the collection of all the functions $\boldsymbol{\gamma}$ of the following form

$$\boldsymbol{\gamma}(\boldsymbol{w})=f\circ g(\boldsymbol{w}),\quad\text{for all }\boldsymbol{w}=(w_1,\ldots,w_d)\in[0,1]^d, \tag{24}$$

where $g:[0,1]^d\to[0,1]^l$ defined as $g(\boldsymbol{w})=(g_1(\boldsymbol{w}),\ldots,g_l(\boldsymbol{w}))$ with

$$g_j(\boldsymbol{w})=\frac{1}{2}\left[\langle a_j,\boldsymbol{w}\rangle+1\right],\ a_j\in\mathcal{C}_d\quad\text{for all }\ j\in\{1,\ldots,l\}$$

and $f\in\mathcal{H}^{\boldsymbol{\alpha}}(L,v_-,v_+)$ mapping $[0,1]^l$ to $[v_-,v_+]\subset I$ with $L\in\mathbb{R}^*_+$, $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_l)\in(\mathbb{R}^*_+)^l$ and $v_-<v_+$. We assume $\boldsymbol{\gamma}^\star\in\mathcal{G}_{[v_-,v_+]}(\boldsymbol{\alpha},L)$ but without knowing the values of $\boldsymbol{\alpha}$ and L.

To approximate the Hölder classes on $[0,1]^l$ with values in $[v_-,v_+]$, we adopt the same strategy by considering the family $\left\{\overline{\mathbf{\Gamma}}^{\mathcal{H},l}_{(\boldsymbol{t},r)},\ (\boldsymbol{t},r)\in(\mathbb{N}^*)^l\times\mathbb{N}\right\}$, where $\overline{\mathbf{\Gamma}}^{\mathcal{H},l}_{(\boldsymbol{t},r)}=\left\{(\boldsymbol{\gamma}\vee v_-)\wedge v_+,\ \boldsymbol{\gamma}\in\overline{\boldsymbol{S}}^{\mathcal{H},l}_{(\boldsymbol{t},r)}\right\}$. Let $[l]=\{1,\ldots,l\}$. For any $r\in\mathbb{N}$ and $\boldsymbol{t}=(t_1,\ldots,t_l)\in(\mathbb{N}^*)^l$, we define the class of functions $\overline{\mathbf{\Gamma}}^M_{(\boldsymbol{t},r)}$ on $\mathscr{W}=[0,1]^d$ as

$$\overline{\mathbf{\Gamma}}^M_{(\boldsymbol{t},r)}=\left\{f\left(g_1(\cdot),\ldots,g_l(\cdot)\right),\ f\in\overline{\mathbf{\Gamma}}^{\mathcal{H},l}_{(\boldsymbol{t},r)},\ g_j=\frac{1}{2}\left[\langle a_j,\cdot\rangle+1\right],\ a_j\in\mathcal{C}_d,\ j\in[l]\right\}.$$

The following result entails that $\overline{\mathbf{\Gamma}}^M_{(\boldsymbol{t},r)}$ is VC-subgraph on $\mathscr{W}$.

Proposition 2. *For any $r \in \mathbb{N}$ and $\boldsymbol{t} = (t_1, \ldots, t_l) \in (\mathbb{N}^*)^l$, the class of functions $\overline{\mathbf{\Gamma}}^M_{(\boldsymbol{t},r)}$ is a VC-subgraph on $\mathscr{W} = [0,1]^d$ with dimension*

$$V^M_{(\boldsymbol{t},r)} \leq 2 + \left[2ld + \left(\prod_{j=1}^{l} t_j \right) (r+1)^l \right] \log_2 \left[4eU \log_2 (2eU) \right], \tag{25}$$

where $U = \sum_{j=1}^{l} t_j + lr + l + 1$.

The proof is postponed to Appendix C. For any $r \in \mathbb{N}$ and $\boldsymbol{t} = (t_1, \ldots, t_l) \in (\mathbb{N}^*)^l$, we take the countable subset $\mathbf{\Gamma}^M_{(\boldsymbol{t},r)}$ of $\overline{\mathbf{\Gamma}}^M_{(\boldsymbol{t},r)}$ defined as

$$\mathbf{\Gamma}^M_{(\boldsymbol{t},r)} = \left\{ f\left(g_1(\cdot), \ldots, g_l(\cdot)\right),\ f \in \mathbf{\Gamma}^{\mathcal{H},l}_{(\boldsymbol{t},r)},\ g_j = \frac{[\langle a_j, \cdot \rangle + 1]}{2},\ a_j \in \mathcal{C}_d \cap \mathbb{Q}^d,\ j \in [l] \right\},$$

where $\mathbf{\Gamma}^{\mathcal{H},l}_{(\boldsymbol{t},r)} = \left\{ (\boldsymbol{\gamma} \vee v_-) \wedge v_+,\ \boldsymbol{\gamma} \in \boldsymbol{S}^{\mathcal{H},l}_{(\boldsymbol{t},r)} \right\}$ with $\boldsymbol{S}^{\mathcal{H},l}_{(\boldsymbol{t},r)}$ the countable subset of $\overline{\boldsymbol{S}}^{\mathcal{H},l}_{(\boldsymbol{t},r)}$ as we introduced in the beginning of this section.

Let $\mathcal{M} = (\mathbb{N}^*)^l \times \mathbb{N}$. For any $r \in \mathbb{N}$ and $\boldsymbol{t} \in (\mathbb{N}^*)^l$, we associate it with the weight

$$\Delta(\boldsymbol{t}, r) = \sum_{j=1}^{l} t_j + r. \tag{26}$$

The following result shows inequality (7) is satisfied with the weights defined by (26).

Lemma 5. *With the weights defined by (26), we have*

$$\sum_{(\boldsymbol{t},r) \in (\mathbb{N}^*)^l \times \mathbb{N}} e^{-\Delta(\boldsymbol{t},r)} \leq \frac{e}{e-1}.$$

The proof is postponed to Appendix B. With Proposition 2 and Lemma 5, we are able to apply the model selection procedure introduced in Section 3.2 and obtain the following.

Corollary 3. *Under Assumption 2, no matter what the distribution of W is, the estimator $\widehat{\boldsymbol{\gamma}}(\boldsymbol{X})$ given by the model selection procedure in Section 3.2 over $\left\{ \mathbf{\Gamma}^M_{(\boldsymbol{t},r)},\ (\boldsymbol{t},r) \in (\mathbb{N}^*)^l \times \mathbb{N} \right\}$ with the weights defined by (26) satisfies for all $\boldsymbol{\alpha} \in (\mathbb{R}^*_+)^l$ and $L > 0$,*

$$\sup_{\boldsymbol{\gamma}^\star \in \mathcal{G}_{[v_-, v_+]}(\boldsymbol{\alpha}, L)} \mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}}) \right] \leq C_{\kappa,l,\boldsymbol{\alpha}} \left(L^{\frac{2l}{2\overline{\alpha}+l}} n^{-\frac{2\overline{\alpha}}{2\overline{\alpha}+l}} + \frac{d}{n} \right) \mathcal{L}_n^2,$$

where $\mathcal{L}_n = \log n \vee \log L^2 \vee 1$ and $C_{\kappa,l,\boldsymbol{\alpha}}$ is a constant depending only on κ, l and $\boldsymbol{\alpha}$.

The result tells that if for some $\boldsymbol{\alpha} \in (\mathbb{R}_+^*)^l$ and $L > 0$, $\boldsymbol{\gamma}^\star \in \mathcal{G}_{[v_-,v_+]}(\boldsymbol{\alpha}, L)$ where the value of l is smaller than d, we mitigate the curse of dimensionality by taking the information that $\boldsymbol{\gamma}^\star$ is a multiple index function. If it is not the case, a bias term will be added into the risk bound in Corollary 3. But as long as the conditional distribution $R^\star$ is not far away from some set of conditional distributions $\left\{R_{\boldsymbol{\gamma}},\ \boldsymbol{\gamma} \in \mathcal{G}_{[v_-,v_+]}(\boldsymbol{\alpha}, L)\right\}$, the performance of our estimator will not deteriorate too much.

When the value of l is large ($l > d$), the multiple index model (24) does not help to circumvent the curse of dimensionality. In this situation, we could assume $\boldsymbol{\gamma}^\star$ has an additive structure, i.e.

$$\boldsymbol{\gamma}^\star(\boldsymbol{w}) = \sum_{j=1}^{l} \boldsymbol{\gamma}_j \left(\frac{\langle a_j, \boldsymbol{w} \rangle + 1}{2} \right), \quad \text{for all } \boldsymbol{w} \in [0,1]^d,$$

where $a_j \in \mathcal{C}_d$. Imposing some smoothness on $\boldsymbol{\gamma}_j$, we can construct models and perform our model selection procedure to mitigate the curse of dimensionality. The construction is similar to a combination of what we have done in Section 5.1 and 5.2.

6. Model selection for neural networks

Throughout this section, we assume the covariates W_i are i.i.d. on $[0,1]^d$ with the common distribution P_W and $R_i^\star = R_{\boldsymbol{\gamma}^\star}$ for all $i \in \{1, \dots, n\}$. The idea in this section is to estimate the regression function $\boldsymbol{\gamma}^\star$ by our model selection procedure based on ReLU neural networks.

We start with setting some notations. We recall the Rectifier Linear Unit (ReLU) activation function $\sigma : \mathbb{R} \to \mathbb{R}$ defined as

$$\sigma(x) = \max(0, x).$$

For any vector $\boldsymbol{x} = (x_1, \dots, x_p)^\top \in \mathbb{R}^p$ with some $p \in \mathbb{N}^*$, by writing $\sigma(\boldsymbol{x})$ we mean the activation function operating component-wise, i.e.

$$\sigma(\boldsymbol{x}) = (\max\{0, x_1\}, \dots, \max\{0, x_d\})^\top.$$

We formulate $\overline{\boldsymbol{S}}_{(L,p)}$ the Multi-Layer Perception (MLP) with width $p \in \mathbb{N}^*$ and depth $L \in \mathbb{N}^*$, which is a collection of functions of the form

$$f : \mathbb{R}^d \to \mathbb{R}, \quad \boldsymbol{w} \mapsto f(\boldsymbol{w}) = M_L \circ \sigma \circ M_{L-1} \circ \cdots \circ \sigma \circ M_0(\boldsymbol{w}), \tag{27}$$

where

$$M_l(\boldsymbol{y}) = A_l(\boldsymbol{y}) + b_l, \quad \text{for } l = 0, \dots, L,$$

A_l is a $p \times p$ weight matrix for $l \in \{1, \dots, L-1\}$, A_0 has size $p \times d$, A_L has size $1 \times p$ and the shift vectors b_l is of size p if $l \in \{0, \dots, L-1\}$, a scalar if $l = L$. All the parameters in weight matrices and shift vectors vary in $\mathbb{R}$. We denote the MLP as $\boldsymbol{S}_{(L,p)}$ when it has the same architecture as $\overline{\boldsymbol{S}}_{(L,p)}$ but all the parameters in weight matrices and shift vectors vary in $\mathbb{Q}$.

Besides learning all the parameters in weight matrices and shift vectors, people also enforce their algorithm on some sparse neural networks depending on the problem they want to solve. Some examples can be found in Section 7.10 of Goodfellow et al. (2016). Another more tuitive example for the sparse setting is the convolutional neural network (CNN) which has been widely used in computer vision, sequence analysis in bioinformatics and natural language processing.

We formulate the sparse ReLU neural networks as follows. For $l \in \{0, \dots, L\}$, we define $\boldsymbol{s}_l$ the indicator vector in which the component is either 0 or 1. The size of the vector $\boldsymbol{s}_l$ equals to the total number of parameters in weight matrix A_l and shift vector b_l. For $l = 0$, $\boldsymbol{s}_0$ is of size $p(d+1)$, for $l \in \{1, \dots, L-1\}$, $\boldsymbol{s}_l$ is of size $p(p+1)$ and for $l = L$, $\boldsymbol{s}_L$ is of size $p+1$. Essentially, indicator vectors $\boldsymbol{s}_l$, $l \in \{0, \dots, L\}$ represent collections of functions based on the structure of neural networks. The last p components in $\boldsymbol{s}_l$, $l \in \{0, \dots, L-1\}$ and the last one in $\boldsymbol{s}_L$ address to the collection of shift vectors b_l. More precisely, for any component in b_l if the corresponding position in $\boldsymbol{s}_l$ is 1, we allow this component in b_l varies in $\mathbb{R}$ otherwise the value of it is fixed at 0. The other components in $\boldsymbol{s}_l$ address to the collection of weight matrices A_l with the same way as we have introduced to b_l after reshaping the matrices one row after another into vectors. To illustrate, we take $p = 2$, $L = 3$ and $l = 1$ as an example. Let $\boldsymbol{s}_1 = (1, 0, 0, 1, 1, 0)^\top$ which is a vector of size 6. As mentioned before, A_1 is a 2×2 matrix which we write as

$$A_1 = \begin{pmatrix} a_1 & a_3 \\ a_4 & a_2 \end{pmatrix}$$

and b_1 is of size 2. The last 2 components in $\boldsymbol{s}_1$ is $(1, 0)^\top$ which entails that the first component in b_1 varies in $\mathbb{R}$ and the second is fixed at 0. We then reshape A_1 one row after another into a vector, namely $(a_1, a_3, a_4, a_2)^\top$. The remaining components of $\boldsymbol{s}_l$ is $(1, 0, 0, 1)^\top$ which entails a_1 and a_2 are allowed varying in $\mathbb{R}$ while $a_3, a_4 = 0$. To conclude, such an indicator vector $\boldsymbol{s}_1$ corresponds to the collection of weight matrices

$$A_1 = \begin{pmatrix} a_1 & 0 \\ 0 & a_2 \end{pmatrix}, \quad \text{with } a_1, a_2 \in \mathbb{R}$$

and shift vectors $b_1 = (b, 0)^\top$ with $b \in \mathbb{R}$.

Given $p, L \in \mathbb{N}^*$ and a joint indicator vector $\boldsymbol{s} = (\boldsymbol{s}_0^\top, \dots, \boldsymbol{s}_L^\top)^\top$, we denote $\overline{\boldsymbol{S}}_{(L,p,\boldsymbol{s})}$ as the corresponding collection of functions on $[0,1]^d$. Similarly, $\boldsymbol{S}_{(L,p,\boldsymbol{s})}$ denotes the class of functions with the same architecture as $\overline{\boldsymbol{S}}_{(L,p,\boldsymbol{s})}$, where the non-zero parameters vary in $\mathbb{Q}$ but not $\mathbb{R}$. Let us remark that given $L \in \mathbb{N}^*$, $p \in \mathbb{N}^*$, $\boldsymbol{s}$ is of size

$$\overline{p} = p^2(L-1) + p(L+d+1) + 1.$$

The following result gives an upper bound of the VC dimension for the class of the functions $\overline{\boldsymbol{S}}_{(L,p,\boldsymbol{s})}$ on $\mathscr{W} = [0,1]^d$.

Proposition 3. *For any $L \in \mathbb{N}^*$, $p \in \mathbb{N}^*$ and $\boldsymbol{s} \in \{0,1\}^{\overline{p}}$, a fixed designed neural network $\overline{\boldsymbol{S}}_{(L,p,\boldsymbol{s})}$ is a VC-subgraph on $\mathscr{W}$ with dimension*

$$V_{(L,p,\boldsymbol{s})} \leq (L+1)(\|\boldsymbol{s}\|_0 + 1) \log_2 \left[2 \left(2e(L+1) \left(\frac{pL}{2} + 1 \right) \right)^2 \right],$$

where $\|\boldsymbol{s}\|_0$ denotes the number of non-zero components in $\boldsymbol{s}$.

The proof is postponed to Appendix C. In particular, when all the components in $\boldsymbol{s}$ are 1, $\overline{\boldsymbol{S}}_{(L,p,\boldsymbol{s})}$ is the Multi-Layer Perception $\overline{\boldsymbol{S}}_{(L,p)}$ and Proposition 3 entails the VC dimension of $\overline{\boldsymbol{S}}_{(L,p)}$ is, up to a constant, bounded by $\overline{p}L \log\left[(L+1)\left(pL/2+1\right)\right]$.

6.1. **The Takagi class of functions.** We provide an example in this subsection where estimation based on ReLU neural networks enjoys a significant advantage.

Let $v_-, v_+ \in \mathbb{R}$ such that $v_- < v_+$ and $[v_-, v_+] \subset I$. For any $t \in (-1,1)$, $l \in \mathbb{N}^*$, $\mathbf{p} = (p_1, p_2) \in \mathbb{N}^* \times \mathbb{N}^*$ and $K \geq 0$, we denote $\mathcal{F}_{[v_-,v_+]}(t, l, \mathbf{p}, K)$ the collection of functions where for all $f \in \mathcal{F}_{[v_-,v_+]}(t, l, \mathbf{p}, K)$, it takes values in $[v_-, v_+] \subset I$ and is of the form

$$f(w) = \sum_{k \in \mathbb{N}^*} t^k g\left(h^{\circ k}(w)\right), \quad \text{for all } w \in [0,1], \tag{28}$$

where $g \in \boldsymbol{S}_{(l,p_1)}$ defined on $[0,1]$, $\|g\|_\infty \leq K$, $h \in \boldsymbol{S}_{(l,p_2)}$ maps $[0,1]$ to $[0,1]$ and $h^{\circ k} = h \circ \cdots \circ h$ denotes the resulted function when h is composed with itself k times. We assume the regression function $\boldsymbol{\gamma}^\star \in \mathcal{F}_{[v_-,v_+]}(t, l, \mathbf{p}, K)$ but without the knowledge of t, l, $\mathbf{p}$ and K. This type of setting provides elementary examples of self similar functions and dynamical systems (see Yamaguti and Hata (1983) for example). It also includes a number of interesting functions belonging to the Takagi class (Daubechies et al. (2019) p.28), which is defined as the collection of all the functions of the form

$$f = \sum_{k \in \mathbb{N}^*} c_k h^{\circ k},$$

where $(c_k)_{k \in \mathbb{N}^*}$ is an absolutely summable sequence of real numbers and h is the hat function defined on $[0,1]$ as

$$h(w) = \begin{cases} 2w & , \quad 0 \leq w \leq \frac{1}{2}, \\ 2(1-w) & , \quad \frac{1}{2} < w \leq 1. \end{cases} \tag{29}$$

Let $\mathcal{M} = \mathbb{N}^* \times \mathbb{N}^*$. The family of models we consider here is given by

$$\left\{ \boldsymbol{\Gamma}_{(L,p)},\ (L,p) \in \mathbb{N}^* \times \mathbb{N}^* \right\},$$

where $\boldsymbol{\Gamma}_{(L,p)} = \left\{ (\boldsymbol{\gamma} \vee v_-) \wedge v_+,\ \boldsymbol{\gamma} \in \boldsymbol{S}_{(L,p)} \right\}$. We note that for each $\boldsymbol{\Gamma}_{(L,p)}$, it is a countable collection of functions on $[0,1]$ and satisfies Assumption 1

with $V_{(L,p)}$, up to a constant, bounded by $\overline{p}L\log\left[(L+1)\left(pL/2+1\right)\right]$. For any $(L,p)\in\mathcal{M}$, we associate it with the weight

$$\Delta(L,p)=L+p. \tag{30}$$

As an immediate consequence, we have $\Sigma=\sum_{(L,p)\in(\mathbb{N}^*)^2}e^{-\Delta(L,p)}\leq 1$ which satisfies the inequality (7). Therefore, we are able to apply the model selection procedure introduced in Section 3.2 and obtain the following result.

Corollary 4. *Under Assumption 2, whatever the distribution of W, the estimator $\widehat{\boldsymbol{\gamma}}(\boldsymbol{X})$ given by the model selection procedure in Section 3.2 over the family $\left\{\boldsymbol{\Gamma}_{(L,p)},\ (L,p)\in\mathbb{N}^*\times\mathbb{N}^*\right\}$ with the weights defined by (30) satisfies for all $t\in(-1,1)$, $l\in\mathbb{N}^*$, $\mathbf{p}\in(\mathbb{N}^*)^2$ and $K\geq 0$*

$$\sup_{\boldsymbol{\gamma}^\star\in\mathcal{F}_{[v_-,v_+]}(t,l,\mathbf{p},K)}\mathbb{E}\left[h^2\left(R_{\boldsymbol{\gamma}^\star},R_{\widehat{\boldsymbol{\gamma}}}\right)\right]\leq C_{\kappa,t,l,\mathbf{p},K}\frac{1}{n}(1+\log n)^4,$$

where $C_{\kappa,t,l,\mathbf{p},K}$ is a constant depending on $\kappa,t,l,\mathbf{p},K$ only.

The risk bound is optimal up to the logarithmic factors since any two probabilities with a Hellinger distance smaller than $\mathcal{O}(1/\sqrt{n})$ are indistinguishable. To comment upon this result further, we consider a specific example of $\boldsymbol{\gamma}^\star$ in Gaussian regression problem with a known variance $\sigma>0$. We parametrize the exponential family $\widetilde{\mathscr{Q}}=\{R_\gamma,\ \gamma\in I\}$ by taking $\gamma=\theta/(2\sqrt{2}\sigma)$, where θ is the mean so that by Proposition 2 of Baraud and Chen (2020), Assumption 2 is satisfied with $\kappa=1$ and $I=\mathbb{R}$. We therefore can take v_- the smallest integer in computer and v_+ the largest so that $[v_-,v_+]\subset I$. Let $\boldsymbol{\gamma}^\star=\sum_{k\in\mathbb{N}^*}2^{-k}h^{\circ k}$ with h defined by (29) be a function belonging to the Takagi class. This corresponds to the situation where g is the identity function on $[0,1]$ so that $K=1$ and $t=1/2$ in the general formalization (28). We also observe that $g\in\boldsymbol{S}_{(1,1)}$ and $h\in\boldsymbol{S}_{(1,2)}$ by rewriting them into the following forms

$$g(x)=\sigma\left(x+0\right),\quad \text{for all } x\in[0,1]$$

and

$$h(w)=\begin{pmatrix}2 & -4\end{pmatrix}\sigma\left\{\begin{pmatrix}1\\1\end{pmatrix}w+\begin{pmatrix}0\\-\frac{1}{2}\end{pmatrix}\right\},\quad \text{for all } w\in[0,1].$$

Therefore, we have $\boldsymbol{\gamma}^\star\in\mathcal{F}_{[v_-,v_+]}(1/2,1,(1,2),1)$. According to Corollary 4, the estimator $\widehat{\boldsymbol{\gamma}}(\boldsymbol{X})$ obtained by the model selection procedure introduced in Section 3.2 based on the fully connected ReLU neural networks converges to $\boldsymbol{\gamma}^\star$ with a rate of order $1/n$ up to logarithmic factors. However, $\boldsymbol{\gamma}^\star$ is nowhere differentiable hence it has very little smoothness in the classical sense. Estimation based on the traditional models will result a miserably slow rate considering the minimax converge rate for an α-smooth function is of order $n^{-2\alpha/(2\alpha+1)}$.

6.2. **Composite Hölder class of functions.** We have seen in the last subsection that the estimator $\widehat{\boldsymbol{\gamma}}(\boldsymbol{X})$ based on MLPs converges to the truth with an optimal rate for some class of functions. In this subsection, we continue to consider the problem of circumventing the curse of dimensionality based on deep ReLU neural networks. A natural structure of the regression function $\boldsymbol{\gamma}^\star$ for neural networks to exhibit advantages could be a composition of several functions which has been considered by Schmidt-Hieber (2020) for Gaussian regression. We shall reconsider it from another point of view where we perform our model selection procedure based on the result of controlling the VC dimension of sparse ReLU neural networks.

Let us introduce notations first. Given $t \in \mathbb{N}^*$ and $\alpha \in \mathbb{R}_+^*$, we define $\mathcal{C}_t^\alpha(D, K)$ an α-Hölder ball with radius K as the collection of functions $f : D \subset \mathbb{R}^t \to \mathbb{R}$ such that

$$\sum_{\substack{\boldsymbol{\beta}=(\beta_1,\dots,\beta_t)\in\mathbb{N}^t \\ \sum_{j=1}^t \beta_j<\alpha}} \|\partial^{\boldsymbol{\beta}} f\|_\infty + \sum_{\substack{\boldsymbol{\beta}\in\mathbb{N}^t \\ \sum_{j=1}^t \beta_j=\lfloor\alpha\rfloor}} \sup_{\substack{\boldsymbol{x},\boldsymbol{y}\in D \\ \boldsymbol{x}\neq\boldsymbol{y}}} \frac{\left|\partial^{\boldsymbol{\beta}} f(\boldsymbol{x}) - \partial^{\boldsymbol{\beta}} f(\boldsymbol{y})\right|}{|\boldsymbol{x}-\boldsymbol{y}|_\infty^{\alpha-\lfloor\alpha\rfloor}} \leq K,$$

where for any $\boldsymbol{\beta} = (\beta_1, \dots, \beta_t) \in \mathbb{N}^t$, $\partial^{\boldsymbol{\beta}} = \partial^{\beta_1} \cdots \partial^{\beta_t}$ and for any $\boldsymbol{x} = (x_1, \dots, x_t) \in \mathbb{R}^t$, $|\boldsymbol{x}|_\infty = \max_{i=1,\dots,t} |x_i|$.

For any $k \in \mathbb{N}^*$, $\mathbf{d} = (d_0, \dots, d_k) \in (\mathbb{N}^*)^{k+1}$, $\mathbf{t} = (t_0, \dots, t_k) \in (\mathbb{N}^*)^{k+1}$, $\boldsymbol{\alpha} = (\alpha_0, \dots, \alpha_k) \in (\mathbb{R}_+^*)^{k+1}$ and $K \geq 0$, we denote $\mathcal{F}_{[v_-,v_+]}(k, \mathbf{d}, \mathbf{t}, \boldsymbol{\alpha}, K)$ the class of functions with values in $[v_-, v_+] \subset I$ as,

$$\mathcal{F}_{[v_-,v_+]}(k, \mathbf{d}, \mathbf{t}, \boldsymbol{\alpha}, K) = \Big\{ f_k \circ \cdots \circ f_0,\ f_i = (f_{ij})_j : [a_i, b_i]^{d_i} \to [a_{i+1}, b_{i+1}]^{d_{i+1}},$$
$$f_{ij} \in \mathcal{C}_{t_i}^{\alpha_i}([a_i, b_i]^{t_i}, K), \text{ for some } |a_i|, |b_i| \leq K \Big\},$$

where $d_{k+1} = 1$. We assume the regression function $\boldsymbol{\gamma}^\star = \boldsymbol{\gamma}_k \circ \cdots \circ \boldsymbol{\gamma}_0 \in \mathcal{F}_{[v_-,v_+]}(k, \mathbf{d}, \mathbf{t}, \boldsymbol{\alpha}, K)$ but without the knowledge of k, $\mathbf{d}$, $\mathbf{t}$, $\boldsymbol{\alpha}$ and K.

To approximate these classes of functions $\mathcal{F}_{[v_-,v_+]}(k, \mathbf{d}, \mathbf{t}, \boldsymbol{\alpha}, K)$, we consider the sparse ReLU neural networks. Recall that for any $(L, p) \in (\mathbb{N}^*)^2$, $\boldsymbol{s} = (\boldsymbol{s}_0^\top, \dots, \boldsymbol{s}_L^\top)^\top \in \{0,1\}^{\overline{p}}$ with $\overline{p} = p^2(L-1) + p(L+d+1) + 1$ indicating the sparsity design of a MLP with architecture (L, p). More precisely, setting $\mathcal{M} = (\mathbb{N}^*)^2 \times \{0,1\}^{\overline{p}}$, we consider the family of models based on sparse ReLU neural networks $\big\{\overline{\boldsymbol{\Gamma}}_{(L,p,\boldsymbol{s})},\ (L, p, \boldsymbol{s}) \in (\mathbb{N}^*)^2 \times \{0,1\}^{\overline{p}}\big\}$, where $\overline{\boldsymbol{\Gamma}}_{(L,p,\boldsymbol{s})} = \big\{(\boldsymbol{\gamma} \vee v_-) \wedge v_+,\ \boldsymbol{\gamma} \in \overline{\boldsymbol{S}}_{(L,p,\boldsymbol{s})}\big\}$. The VC dimension $V_{(L,p,\boldsymbol{s})}$ of each $\overline{\boldsymbol{\Gamma}}_{(L,p,\boldsymbol{s})}$ is bounded by Proposition 3. For each $(L, p, \boldsymbol{s}) \in \mathcal{M}$, we take the countable subset $\boldsymbol{\Gamma}_{(L,p,\boldsymbol{s})}$ of $\overline{\boldsymbol{\Gamma}}_{(L,p,\boldsymbol{s})}$ as $\boldsymbol{\Gamma}_{(L,p,\boldsymbol{s})} = \big\{(\boldsymbol{\gamma} \vee v_-) \wedge v_+,\ \boldsymbol{\gamma} \in \boldsymbol{S}_{(L,p,\boldsymbol{s})}\big\}$.

For each $(L, p, \boldsymbol{s}) \in (\mathbb{N}^*)^2 \times \{0,1\}^{\overline{p}}$, we associate it with the weight

$$\Delta(L, p, \boldsymbol{s}) = \begin{cases} \|\boldsymbol{s}\|_0 \log\left(\dfrac{2e\overline{p}}{\|\boldsymbol{s}\|_0}\right) + p + L & ,\quad \|\boldsymbol{s}\|_0 \neq 0, \\ p + L & ,\quad \|\boldsymbol{s}\|_0 = 0. \end{cases} \tag{31}$$

The following result shows (7) is satisfied with the associated weights defined by (31).

Lemma 6. *For any $L \in \mathbb{N}^*$, $p \in \mathbb{N}^*$ and $\boldsymbol{s} = (\boldsymbol{s}_0^\top, \dots, \boldsymbol{s}_L^\top)^\top \in \{0,1\}^{\overline{p}}$, we define $\Delta(L, p, \boldsymbol{s})$ by (31). Then,*

$$\sum_{(L,p,\boldsymbol{s}) \in (\mathbb{N}^*)^2 \times \{0,1\}^{\overline{p}}} e^{-\Delta(L,p,\boldsymbol{s})} \leq 2.$$

For any $\boldsymbol{\alpha} = (\alpha_0, \dots, \alpha_k) \in (\mathbb{R}_+^*)^{k+1}$, we define the effective smoothness indices by $\alpha_i' = \alpha_i \prod_{l=i+1}^{k} (\alpha_l \wedge 1)$ for all $i \in \{0, \dots, k-1\}$ and $\alpha_k' = \alpha_k$. We denote $\phi_n = \max_{i=0,\dots,k} n^{-2\alpha_i'/(2\alpha_i'+t_i)}$. Combining the result of Lemma 6 and Proposition 3, we are now able to apply the model selection procedure in Section 3.2. The following result entails the estimator $\widehat{\boldsymbol{\gamma}}(\boldsymbol{X})$ converges to $\boldsymbol{\gamma}^\star$ with a rate of order ϕ_n up to logarithm factors with respect to the distance $d(\boldsymbol{\gamma}^\star, \widehat{\boldsymbol{\gamma}}) = h^2(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}})$.

Corollary 5. *Under Assumption 2, whatever the distribution of W, the estimator $\widehat{\boldsymbol{\gamma}}(\boldsymbol{X})$ given by the model selection procedure in Section 3.2 over the family $\left\{\boldsymbol{\Gamma}_{(L,p,\boldsymbol{s})},\ (L,p,\boldsymbol{s}) \in (\mathbb{N}^*)^2 \times \{0,1\}^{\overline{p}}\right\}$ with the weights defined by (31) satisfies with a sufficiently large n, for all $k \in \mathbb{N}^*$, $K \geq 0$, $\mathbf{d} \in (\mathbb{N}^*)^{k+1}$, $\mathbf{t} \in (\mathbb{N}^*)^{k+1}$ with $t_j \leq d_j$ for $j \in \{0, \dots, k\}$ and $\boldsymbol{\alpha} \in (\mathbb{R}_+^*)^{k+1}$,*

$$\sup_{\boldsymbol{\gamma}^\star \in \mathcal{F}_{[v_-,v_+]}(k,\mathbf{d},\mathbf{t},\boldsymbol{\alpha},K)} \mathbb{E}\left[h^2\left(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}}\right)\right] \leq C_{\kappa,k,\mathbf{d},\mathbf{t},\boldsymbol{\alpha},K} \phi_n \log^4 n, \tag{32}$$

where $C_{\kappa,k,\mathbf{d},\mathbf{t},\boldsymbol{\alpha},K}$ is a constant depending on $\kappa, k, \mathbf{d}, \mathbf{t}, \boldsymbol{\alpha}, K$ only.

By Corollary 5, we provide a theoretical guarantee for an alternative estimation procedure based on sparse ReLU neural networks besides maximum likelihood estimation (MLE) discussed in Schmidt-Hieber (2020) for the Gaussian regression. Our procedure is, however, designed to handle the regression problems in exponential families and not only restricted to the Gaussian case. It also endows the estimator an additional robust property compared to MLE. When there is a misspecification or data contamination, as long as the bias remains small compared to the right hand side of (32), the behaviour of our estimator will be of the same order as the model is exact.

7. Variable selection in exponential families

In this section, we propose to handle variable selection problem in exponential families by model selection. The statistical setting is stated as follows. Assuming that W_i are i.i.d. on $\mathscr{W} \subset \mathbb{R}^p$ and for each $i \in \{1, \dots, n\}$, we observe $X_i = (W_i^{(1)}, \dots, W_i^{(p)}, Y_i)$ where $W_i^{(j)}$ represents the observation of the explanatory variable $W^{(j)}$ in the i-th experiment. The value p stands for the number of the explanatory variables. This number may be large,

possibly larger than n. The exponential family $\widetilde{\mathscr{Q}} = \{R_\gamma = r_\gamma \cdot \nu,\ \gamma \in I\}$ is parametrized in its natural form, i.e. for all $y \in \mathscr{Y}$, $\gamma \in I$,

$$r_\gamma(y) = e^{\gamma T(y) - B(\gamma)},$$

which is the particular situation when taking u as the identity function in (2). We assume that there exists an unknown function $\boldsymbol{\gamma}^\star$ on $\mathscr{W}$ taking values in $[v_-, v_+] \subset I$ with $v_- < v_+$ as a linear combination of some subset of the p explanatory variables, namely

$$\boldsymbol{\gamma}^\star(\boldsymbol{w}) = \sum_{j=1}^{p} \gamma_j^\star w^{(j)} \quad \text{for all } \boldsymbol{w} = (w^{(1)}, \ldots, w^{(p)}) \in \mathscr{W},$$

with $\gamma_j^\star \in \mathbb{R}$, such that the conditional distribution of Y_i given W_i belongs to a natural exponential family with natural parameter $\boldsymbol{\gamma}^\star(W_i)$, i.e. $R_{\boldsymbol{\gamma}^\star(W_i)}$. Variable selection problem attributes to estimate this unknown $\boldsymbol{\gamma}^\star$ together with selecting the most significant explanatory variables among the p possible ones.

We set $\Omega = \{1, \ldots, p\}$ and $\mathcal{M} = \mathcal{P}(\Omega)$. For any subset $m \in \mathcal{M}$, we define $\overline{\boldsymbol{S}}_m$ as the collection of functions $\boldsymbol{\gamma}$ on $\mathscr{W}$ of the form

$$\boldsymbol{\gamma}(\boldsymbol{w}) = \sum_{j=1}^{p} \gamma_j w^{(j)} \quad \text{for all } \boldsymbol{w} \in \mathscr{W}, \tag{33}$$

where the coordinates of $\widetilde{\gamma} = (\gamma_1, \ldots, \gamma_p) \in \mathbb{R}^p$ are all zeros except for those indices $j \in m$. By convention, $\overline{\boldsymbol{S}}_m = \{0\}$ if $m = \varnothing$. We define $\boldsymbol{S}_m$ as the collection of functions of the form given by (33) with a restriction to the rational combinations, i.e. for any $\boldsymbol{\gamma} \in \boldsymbol{S}_m$, $\widetilde{\gamma} = (\gamma_1, \ldots, \gamma_p) \in \mathbb{Q}^p$. For each $m \in \mathcal{M}$, let us define $\overline{\boldsymbol{\Gamma}}_m = \left\{(\boldsymbol{\gamma} \vee v_-) \wedge v_+,\ \boldsymbol{\gamma} \in \overline{\boldsymbol{S}}_m\right\}$ and $\boldsymbol{\Gamma}_m = \{(\boldsymbol{\gamma} \vee v_-) \wedge v_+,\ \boldsymbol{\gamma} \in \boldsymbol{S}_m\}$. With the fact that $\mathbb{Q}$ is dense in $\mathbb{R}$, $\boldsymbol{S}_m$ is dense in $\overline{\boldsymbol{S}}_m$ for the topology of pointwise convergence. One can observe that such dense property also holds for each $\boldsymbol{\Gamma}_m$ in $\overline{\boldsymbol{\Gamma}}_m$, $m \in \mathcal{M}$.

We define $\mathcal{M}_o = \{m_d = \{1, \ldots, d\},\ 1 \le d \le p\} \cup \varnothing$. For each $m \in \mathcal{M}$, we associate it with the weight

$$\Delta(m) = \begin{cases} 2\log(1 + |m|) & ,\quad m \in \mathcal{M}_o, \\ |m| \log\left(\dfrac{2ep}{|m|}\right) & ,\quad m \in \mathcal{M} \backslash \mathcal{M}_o. \end{cases} \tag{34}$$

The following result shows with the weights defined by (34), inequality (7) is satisfied.

Lemma 7. *Let $\mathcal{M} = \mathcal{P}(\Omega)$. For any $m \in \mathcal{M}$, the weight is defined by (34). Then $\Sigma = \sum_{m \in \mathcal{M}} e^{-\Delta(m)} \le 1 + \pi^2/6$.*

Moreover, for any $m \in \mathcal{M}$, $\overline{\boldsymbol{S}}_m$ defined by (33) is a $|m|$-dimensional vector space. As an immediate consequence, $\overline{\boldsymbol{\Gamma}}_m$ is VC-subgraph on $\mathscr{W}$ with dimension not larger than $|m| + 1$ which satisfies the Assumption 1 with

$V_m = |m| + 1$. We are now able to apply the model selection procedure presented in Section 3.2 and obtain the following result.

Corollary 6. *For all $m \in \mathcal{M}$, let $\overline{\mathscr{Q}}_m = \left\{R_{\boldsymbol{\gamma}},\ \boldsymbol{\gamma} \in \overline{\boldsymbol{\Gamma}}_m\right\}$. Whatever the distribution of W, the estimator $R_{\widehat{\boldsymbol{\gamma}}}$ given by the model selection procedure in Section 3.2 over $\{\boldsymbol{\Gamma}_m,\ m \in \mathcal{M}\}$ associated with the weight defined by (34) satisfies*

$$\mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}})\right] \leq 1.95 \times 10^7 (\mathscr{B}_o \wedge \mathscr{B}_c), \tag{35}$$

where

$$\mathscr{B}_o = \inf_{m \in \mathcal{M}_o} \left\{h^2(R_{\boldsymbol{\gamma}^\star}, \overline{\mathscr{Q}}_m) + \frac{|m|+1}{n}\left[1 + \log_+\left(\frac{n}{|m|+1}\right)\right]\right\}$$

and

$$\mathscr{B}_c = \inf_{m \in \mathcal{M}} \left\{h^2(R_{\boldsymbol{\gamma}^\star}, \overline{\mathscr{Q}}_m) + \frac{|m|+1}{n}\left[1 + \log\left[\frac{(2p) \vee n}{|m|+1}\right]\right]\right\}.$$

The proof of Corollary 6 is postponed to Appendix A. Let us remark a little bit here for the strategy of assigning weights which is different with the typical choice, where for each $m \in \mathcal{M}$,

$$\Delta(m) = \begin{cases} |m| \log\left(\dfrac{2ep}{|m|}\right) & , \quad m \neq \varnothing, \\ 0 & , \quad m = \varnothing. \end{cases}$$

With the typical choice of the associated weights, one can derive a risk bound $\mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}})\right] \leq 1.95 \times 10^7 \mathscr{B}_c$. Comparing this result with the one given in (35), we note that (35) improves it by a $\log(p)$ term whenever the minimizer $m^\star \in \mathcal{M}$ in the right hand side of (35) does belong to $\mathcal{M}_o$.

Appendix A. Proofs of the main theorem and its corollaries

A.1 Proof of Theorem 1. Before starting to prove the main theorem, let us introduce some notations and facts for later use. For all $i \in \{1, \ldots, n\}$, let $P_i^\star$ be the true distribution of $X_i = (W_i, Y_i)$ and $\mathbf{P}^\star = \otimes_{i=1}^n P_i^\star$ be the true joint distribution of the observed data $\boldsymbol{X} = (X_1, \ldots, X_n)$. We denote $\mathbf{P}_{\boldsymbol{\gamma}} = \otimes_{i=1}^n P_{i,\boldsymbol{\gamma}}$ as the distribution of independent random variables $(W_1, Y_1), \ldots, (W_n, Y_n)$ for which the conditional distribution of Y_i given W_i is given by $R_{\boldsymbol{\gamma}(W_i)} \in \widetilde{\mathscr{Q}}$ for each i. With the equalities $P_i^\star = R_i^\star \cdot P_{W_i}$, $P_{i,\boldsymbol{\gamma}} = R_{\boldsymbol{\gamma}} \cdot P_{W_i}$, we have

$$h^2(P_i^\star, P_{i,\boldsymbol{\gamma}}) = \int_{\mathscr{W}} h^2\left(R_i^\star(w), R_{\boldsymbol{\gamma}(w)}\right) dP_{W_i}(w).$$

If we define the pseudo Hellinger distance $\mathbf{h}$ between two probabilities $\mathbf{P} = \otimes_{i=1}^n P_i$ and $\mathbf{P}' = \otimes_{i=1}^n P_i'$ by

$$\mathbf{h}^2(\mathbf{P}, \mathbf{P}') = \sum_{i=1}^n h^2\left(P_i, P_i'\right),$$

for any $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}$, we have

$$\begin{aligned}\mathbf{h}^2(\mathbf{R}^\star, \mathbf{R}_{\boldsymbol{\gamma}}) &= \sum_{i=1}^n \int_{\mathscr{W}} h^2\left(R_i^\star(w), R_{\boldsymbol{\gamma}(w)}\right) dP_{W_i}(w) \\ &= \sum_{i=1}^n h^2\left(P_i^\star, P_{i,\boldsymbol{\gamma}}\right) = \mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\boldsymbol{\gamma}}). \end{aligned} \tag{36}$$

We set $\boldsymbol{\tau} = \bigotimes_{i=1}^n \tau_i$ with $\tau_i = P_{W_i} \otimes \nu$ for all $i \in \{1, \ldots, n\}$. For all $m \in \mathcal{M}$, we denote by $\boldsymbol{\mathcal{R}}_m$ the following families of densities (with respect to $\boldsymbol{\tau}$) on $\mathscr{X}^n = (\mathscr{W} \times \mathscr{Y})^n$

$$\boldsymbol{\mathcal{R}}_m = \{\mathbf{r}_{\boldsymbol{\gamma}} : \mathbf{x} = (x_1, \ldots, x_n) \mapsto r_{\boldsymbol{\gamma}(w_1)}(y_1) \ldots r_{\boldsymbol{\gamma}(w_n)}(y_n),\ \boldsymbol{\gamma} \in \boldsymbol{\Gamma}_m\}$$

and by $\boldsymbol{\mathscr{P}}_m$ the corresponding ρ-model, i.e. the finite or countable set of probabilities $\{\mathbf{P} = \mathbf{r}_{\boldsymbol{\gamma}} \cdot \boldsymbol{\tau},\ \boldsymbol{\gamma} \in \boldsymbol{\Gamma}_m\}$ with the representation $(\boldsymbol{\tau}, \boldsymbol{\mathcal{R}}_m)$.

Proposition 4. *Under Assumption 1, for any $m \in \mathcal{M}$, the class of functions $\mathcal{R}_m = \{r_{\boldsymbol{\gamma}} :\ (w, y) \mapsto r_{\boldsymbol{\gamma}(w)}(y),\ \boldsymbol{\gamma} \in \overline{\boldsymbol{\Gamma}}_m\}$ on $\mathscr{X} = \mathscr{W} \times \mathscr{Y}$ is VC-subgragh with dimension not larger than $9.41V_m$.*

Proof. For any $m \in \mathcal{M}$, reparametrizing the exponential family in its natural form, we obtain

$$\mathcal{R}_m = \{q_{\boldsymbol{\theta}} :\ (w, y) \mapsto e^{T(y)\boldsymbol{\theta}(w) - A(\boldsymbol{\theta}(w))},\ \boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}}_m\},$$

where $A(\theta) = \log\left[\int_{\mathscr{Y}} \exp(\theta T(y)) d\nu(y)\right]$ and $\overline{\boldsymbol{\Theta}}_m = \{\boldsymbol{\theta} = u \circ \boldsymbol{\gamma},\ \boldsymbol{\gamma} \in \overline{\boldsymbol{\Gamma}}_m\}$. By Proposition 42-(ii) of Baraud et al. (2017), VC-subgraph is preserved by composition with a monotone function. Therefore, under Assumption 1, $\overline{\boldsymbol{\Theta}}_m$ is also VC-subgraph on $\mathscr{W}$ with dimension not larger than $V_m \geq 1$. Applying Proposition 5 of Baraud and Chen (2020) with $\mathcal{P} = \mathcal{R}_m$ for each $m \in \mathcal{M}$, we can conclude. □

Let us remark that the function ψ defined by (5) in the present paper satisfies the Assumption 2 in Baraud and Birgé (2018) with $a_0 = 4$, $a_1 = 3/8$, $a_2^2 = 3\sqrt{2}$ and for any ρ-model $\boldsymbol{\mathscr{P}}_m$, we follow the definition of ρ-dimension function $D^{\boldsymbol{\mathscr{P}}_m}$ of $\boldsymbol{\mathscr{P}}_m$ given by (15) in Baraud and Birgé (2018). The next result provides an upper bound for $D^{\boldsymbol{\mathscr{P}}_m}$.

Proposition 5. *Under Assumption 1, for any $m \in \mathcal{M}$, for all product probabilities $\mathbf{P}^\star$ and $\overline{\mathbf{P}} = \otimes_{i=1}^n \overline{P}_i$ on $(\mathscr{X}^n, \mathcal{X}^n)$ with $\overline{P}_i = \overline{p} \cdot \tau_i$ for all $i \in \{1, \ldots, n\}$,*

$$D^{\boldsymbol{\mathscr{P}}_m}(\mathbf{P}^\star, \overline{\mathbf{P}}) \leq 10^3 V_m \left[9.11 + \log_+\left(\frac{n}{V_m}\right)\right].$$

Proof. The proof is basically similar to the proof of Proposition 6 in Baraud and Chen (2020) except a modification of the class $\mathscr{F}_y$. More precisely, for

any $y > 0$, we define

$$\mathscr{F}_y = \left\{ \psi\left(\sqrt{\frac{r_{\boldsymbol{\gamma}}}{\overline{p}}}\right) \middle| \ \boldsymbol{\gamma} \in \boldsymbol{\Gamma}_m,\ \mathbf{h}^2(\mathbf{P}^\star, \mathbf{r}_{\boldsymbol{\gamma}} \cdot \boldsymbol{\tau}) + \mathbf{h}^2(\mathbf{P}^\star, \overline{\mathbf{P}}) < y^2 \right\}.$$

Then combining Proposition 4, the conclusion is easy to obtain by following the proof of Proposition 6 in Baraud and Chen (2020). □

Now we turn to prove Theorem 1. It follows by Proposition 5 taking $\overline{\mathbf{P}} = \mathbf{P}_{\boldsymbol{\gamma}}$ that for any $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}$,

$$D^{\mathscr{P}_m}(\mathbf{P}^\star, \mathbf{P}_{\boldsymbol{\gamma}}) \leq 10^3 V_m \left[9.11 + \log_+\left(\frac{n}{V_m}\right)\right] = D_n(m),$$

which satisfies (22) of Baraud and Birgé (2018) with $K = 0$. Applying Theorem 2 of Baraud and Birgé (2018) over the collection of ρ-models $\{\mathscr{P}_m,\ m \in \mathcal{M}\}$ with $\kappa_1 = 0$, we obtain for any arbitrary $\mathbf{P}^\star$, the ρ-estimator $\mathbf{P}_{\widehat{\gamma}}$ satisfies, for all $\xi > 0$ with a probability at least $1 - \Sigma e^{-\xi}$,

$$\mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\widehat{\gamma}}) \leq \inf_{m \in \mathcal{M}} \left[c_1 \mathbf{h}^2(\mathbf{P}^\star, \mathscr{P}_m) + c_2\left(\Xi(m) + 1.49 + \xi\right)\right], \tag{37}$$

where $c_1 = 149.8$ and $c_2 = 5013.2$. The constant Σ in front of $e^{-\xi}$ just due to in Theorem 2 of Baraud and Birgé (2018) they assumed $\Sigma \leq 1$ for the sake of simplicity. One can refer to their proof of Theorem 2 for understanding the role Σ plays. The conclusion finally follows from the equalities $\mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\widehat{\gamma}}) = \mathbf{h}^2(\mathbf{R}^\star, \mathbf{R}_{\widehat{\gamma}})$ and $\mathbf{h}^2(\mathbf{P}^\star, \mathscr{P}_m) = \mathbf{h}^2(\mathbf{R}^\star, \mathscr{Q}_m)$, for all $m \in \mathcal{M}$.

A.2 Proof of Corollary 1. We first present the following approximation result which is an immediate consequence combining Theorem 1 and Proposition 2 of Akakpo (2012).

Proposition 6. *Let $r \in \mathbb{N}$, $R \in \mathbb{R}_+^*$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) \in \prod_{j=1}^d (0, r+1)$, $p > 0$ and $1 \leq p' \leq \infty$ such that*

$$\frac{\overline{\alpha}}{d} > \left(\frac{1}{p} - \frac{1}{p'}\right)_+.$$

For all $f \in B_{p,q}^{\boldsymbol{\alpha}}\left([0,1]^d, R\right)$ and all $l \in \mathbb{N}$, there exists a partition $\pi(l) \in \cup_{\boldsymbol{s} \in \mathbb{N}^d} M_{\boldsymbol{s}}^{\mathcal{B},d}$ of $[0,1]^d$ containing only hyperrectangles such that

$$|\pi(l)| \leq C_{d,\boldsymbol{\alpha},p} 2^{ld}$$

and

$$\inf_{\widetilde{f} \in \overline{\boldsymbol{S}}_{(\pi(l),r)}^{\mathcal{B},d}} \|f - \widetilde{f}\|_{p'} \leq C_{d,r,\boldsymbol{\alpha},p,p'} R 2^{-l\overline{\alpha}}, \tag{38}$$

where $q = \infty$ if $0 < p \leq 1$ or $p \geq 2$ and $q = p$ if $1 < p < 2$, $C_{d,r,\boldsymbol{\alpha},p,p'}$ is a constant depending only on $d, r, \boldsymbol{\alpha}, p, p'$.

Now we turn to prove Corollary 1. Under Assumption 2, applying (13), Lemma 1 and (2), we derive no matter what the distribution of W is, for all $R \in \mathbb{R}_+^*$, $p > 0$ and $\boldsymbol{\alpha} \in (\mathbb{R}_+^*)^d$ such that $\overline{\alpha}/d > 1/p$, any $\boldsymbol{\gamma}^\star \in B_{p,q}^{\boldsymbol{\alpha}}(R, v_-, v_+)$

$$\begin{aligned}
&\mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}})\right] \\
\leq & c_2\left(c_3 + \frac{e}{e-1}\right) \inf_{(\boldsymbol{s},r)\in\mathcal{M}} \left[h^2(R_{\boldsymbol{\gamma}^\star}, \overline{\mathscr{Q}}_{(\boldsymbol{s},r)}^d) + \frac{\Delta(\boldsymbol{s},r)}{n} + \frac{V_{(\boldsymbol{s},r)}}{n}(1+\log n)\right] \\
& \leq C_\kappa \inf_{(\boldsymbol{s},r)\in\mathcal{M}} \left[\inf_{\overline{\boldsymbol{\gamma}}\in\overline{\boldsymbol{\Gamma}}_{(\boldsymbol{s},r)}^{\mathcal{B},d}} \|\boldsymbol{\gamma}^\star - \overline{\boldsymbol{\gamma}}\|_{2,P_W}^2 + \frac{\Delta(\boldsymbol{s},r)}{n} + \frac{V_{(\boldsymbol{s},r)}}{n}(1+\log n)\right],
\end{aligned} \tag{39}$$

where $\overline{\mathscr{Q}}_{(\boldsymbol{s},r)}^d = \left\{R_{\boldsymbol{\gamma}},\ \boldsymbol{\gamma} \in \overline{\boldsymbol{\Gamma}}_{(\boldsymbol{s},r)}^{\mathcal{B},d}\right\}$ and C_κ is a constant depending on κ only. We then apply Proposition 6 by taking $r = \left\lfloor \sup_{j=1,\ldots,d} \alpha_j \right\rfloor \in \mathbb{N}$, $p' = \infty$ and obtain that for all $l \in \mathbb{N}$, there exists a partition $\pi(l) \in M^{\mathcal{B},d}$ such that

$$\Delta(\pi(l), r) = \log(8d)|\pi(l)| + r \leq C_{d,\boldsymbol{\alpha},p} 2^{ld}, \tag{40}$$

$$V_{(\pi(l),r)} = (r+1)^d |\pi(l)| + 1 \leq C_{d,\boldsymbol{\alpha},p} 2^{ld}, \tag{41}$$

$$\begin{aligned}
\inf_{\overline{\boldsymbol{\gamma}}\in\overline{\boldsymbol{\Gamma}}_{(\pi(l),r)}^{\mathcal{B},d}} \|\boldsymbol{\gamma}^\star - \overline{\boldsymbol{\gamma}}\|_{2,P_W}^2 &= \inf_{\overline{\boldsymbol{\gamma}}\in\overline{\boldsymbol{\Gamma}}_{(\pi(l),r)}^{\mathcal{B},d}} \int_{\mathscr{W}} |\boldsymbol{\gamma}^\star(w) - \overline{\boldsymbol{\gamma}}(w)|^2 dP_W(w) \\
&\leq \inf_{\overline{\boldsymbol{\gamma}}\in\overline{\boldsymbol{\Gamma}}_{(\pi(l),r)}^{\mathcal{B},d}} \|\boldsymbol{\gamma}^\star - \overline{\boldsymbol{\gamma}}\|_\infty^2 \\
&\leq \inf_{\overline{\boldsymbol{\gamma}}\in\overline{\boldsymbol{S}}_{(\pi(l),r)}^{\mathcal{B},d}} \|\boldsymbol{\gamma}^\star - \overline{\boldsymbol{\gamma}}\|_\infty^2 \\
&\leq C_{d,\boldsymbol{\alpha},p} R^2 2^{-2l\overline{\alpha}}.
\end{aligned} \tag{42}$$

Plugging (40), (41) and (42) into (39), we derive

$$\begin{aligned}
&\mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}})\right] \\
\leq & C_\kappa \inf_{l\in\mathbb{N}} \left[\inf_{\overline{\boldsymbol{\gamma}}\in\overline{\boldsymbol{\Gamma}}_{(\pi(l),r)}^{\mathcal{B},d}} \|\boldsymbol{\gamma}^\star - \overline{\boldsymbol{\gamma}}\|_{2,P_W}^2 + \frac{\Delta(\pi(l),r)}{n} + \frac{V_{(\pi(l),r)}}{n}(1+\log n)\right] \\
\leq & C_{\kappa,d,\boldsymbol{\alpha},p} \inf_{l\in\mathbb{N}} \left(R^2 2^{-2l\overline{\alpha}} + \frac{2^{ld}}{n}\right)(1+\log n),
\end{aligned} \tag{43}$$

where $C_{\kappa,d,\boldsymbol{\alpha},p}$ is a constant depending on $\kappa, d, \boldsymbol{\alpha}, p$ only. To conclude, we need to minimize the right hand side of (43). If $nR^2 < 1$, we take $l = 0$ so that

$$R^2 2^{-2l\overline{\alpha}} + \frac{2^{ld}}{n} = R^2 + \frac{1}{n} < \frac{2}{n}. \tag{44}$$

Otherwise, we take l as the largest natural number such that $2^{ld}/n \leq R^2 2^{-2l\overline{\alpha}}$ which is well defined since $nR^2 \geq 1$. With this choice of l,

$$R^2 2^{-2l\overline{\alpha}} + \frac{2^{ld}}{n} \leq 2R^2 2^{-2l\overline{\alpha}} \leq C_{\boldsymbol{\alpha}} R^{\frac{2d}{d+2\overline{\alpha}}} n^{-\frac{2\overline{\alpha}}{d+2\overline{\alpha}}}, \tag{45}$$

where $C_{\boldsymbol{\alpha}}$ is a constant depending only on $\boldsymbol{\alpha}$. Combining (43), (44) and (45), we obtain

$$\mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}})\right] \leq C_{\kappa,d,\boldsymbol{\alpha},p}\left(R^{\frac{2d}{d+2\overline{\alpha}}} n^{-\frac{2\overline{\alpha}}{d+2\overline{\alpha}}} + \frac{1}{n}\right)(1+\log n)\,.$$

We conclude by taking the supremum over the set $B^{\boldsymbol{\alpha}}_{p,q}(R, v_-, v_+)$.

A.3 Proof of Corollary 2.

Lemma 8. *For any $k \in \mathbb{N}^*$, $x_1, \ldots, x_k \geq 0$ and $\alpha \in (0,1]$, $\left(\sum_{i=1}^k x_i\right)^\alpha \leq \sum_{i=1}^k x_i^\alpha$.*

Proof. In fact, it is enough to prove when $k = 2$, i.e. $(x_1 + x_2)^\alpha \leq x_1^\alpha + x_2^\alpha$. If at least one of x_1 and x_2 is equal to zero, then the conclusion is trivial. So we suppose $x_1, x_2 > 0$. The function $f(x) = x^\alpha$ is concave on $(0, +\infty)$ since its second derivative $f''(x) = \alpha(\alpha-1)x^{\alpha-2}$ is always negative for all $x \in (0, +\infty)$. By the definition of the concave function, for any $\lambda \in [0,1]$,

$$(\lambda x)^\alpha = [\lambda x + (1-\lambda)0]^\alpha \geq \lambda x^\alpha.$$

Therefore, for any $x_1, x_2 > 0$

$$\begin{aligned} x_1^\alpha + x_2^\alpha &= \left[\frac{x_1}{x_1+x_2}(x_1+x_2)\right]^\alpha + \left[\frac{x_2}{x_1+x_2}(x_1+x_2)\right]^\alpha \\ &\geq \frac{x_1}{x_1+x_2}(x_1+x_2)^\alpha + \frac{x_2}{x_1+x_2}(x_1+x_2)^\alpha \\ &= (x_1+x_2)^\alpha. \end{aligned}$$

□

We then introduce a result given by Lemma 4 of Baraud and Birgé (2014) which we will use later in the proof.

Lemma 9. *Let $(A, \mathcal{A}, \mu)$ be some probability space and u some nondecreasing and nonnegative concave function on $[0, +\infty)$ such that $u(0) = 0$. For all $k \in [1, +\infty]$ and $h \in \mathbb{L}_k(A, \mu)$,*

$$\|u(|h|)\|_{k,\mu} \leq 2^{1/k} u(\|h\|_{k,\mu}),$$

with the convention $2^{1/\infty} = 1$.

Finally, we introduce the following approximation result which is obtained by combining Corollary 3.1 of Dahmen et al. (1980) and Schumaker (1981) (13.62 p.517). It also appeared in the proof of Proposition 5 in Barron et al. (1999) (4.25 p.347).

Proposition 7. *For a given $k \in \mathbb{N}^*$, let $r \in \mathbb{N}$ such that $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k) \in \prod_{j=1}^k (0, r+1)$. For all $f \in \mathcal{H}^{\boldsymbol{\alpha}}([0,1]^k, L)$ and all $\boldsymbol{t} = (t_1, \ldots, t_k) \in (\mathbb{N}^*)^k$, we have*

$$\inf_{\widetilde{f} \in \overline{\boldsymbol{S}}^{\mathcal{H},k}_{(\boldsymbol{t},r)}} \|f - \widetilde{f}\|_\infty \leq C_{k,r} L \sum_{j=1}^k t_j{}^{-\alpha_j}, \tag{46}$$

where $C_{k,r}$ is a constant depending on k and r.

Now we turn to prove Corollary 2. First, we note that for any function $\boldsymbol{\gamma}^\star = \boldsymbol{\gamma}\left(\sum_{j=1}^d \boldsymbol{\gamma}_j'\right) \in \mathcal{F}_{[v_-,v_+]}(\alpha, \boldsymbol{\beta}, \mathbf{p}, L, \mathbf{R})$ and any $[f\left[(g \vee 0) \wedge 1\right] \vee v_-] \wedge v_+$, where $f \in \overline{\boldsymbol{S}}^{\mathcal{H},1}_{(t,r)}$, $g(\boldsymbol{w}) = \sum_{j=1}^d g_j(w_j)$, $g_j \in \overline{\boldsymbol{S}}_{(\pi_j,r)}$, $(\boldsymbol{\pi}, t, r) \in (M^{\mathcal{B},1})^d \times \mathbb{N}^* \times \mathbb{N}$, with the fact that $\boldsymbol{\gamma} \in \mathcal{H}^\alpha(L, v_-, v_+)$ and $\boldsymbol{\gamma}_j'$ taking values in $[0, 1/d]$ for all $j \in \{1, \ldots, d\}$, we have

$$\begin{aligned}
&\sup_{\boldsymbol{w} \in [0,1]^d} \left| \boldsymbol{\gamma}\left(\sum_{j=1}^d \boldsymbol{\gamma}_j'(w_j)\right) - \left[f\left[\left(\left(\sum_{j=1}^d g_j(w_j)\right) \vee 0\right) \wedge 1\right] \vee v_- \right] \wedge v_+ \right| \\
\leq &\sup_{\boldsymbol{w} \in [0,1]^d} \left| \boldsymbol{\gamma}\left(\sum_{j=1}^d \boldsymbol{\gamma}_j'(w_j)\right) - f\left[\left(\left(\sum_{j=1}^d g_j(w_j)\right) \vee 0\right) \wedge 1\right] \right| \\
\leq &\sup_{\boldsymbol{w} \in [0,1]^d} \left| \boldsymbol{\gamma}\left(\sum_{j=1}^d \boldsymbol{\gamma}_j'(w_j)\right) - \boldsymbol{\gamma}\left[\left(\left(\sum_{j=1}^d g_j(w_j)\right) \vee 0\right) \wedge 1\right] \right| + \|\boldsymbol{\gamma} - f\|_\infty \\
\leq &L \sup_{\boldsymbol{w} \in [0,1]^d} \left| \left(\sum_{j=1}^d \boldsymbol{\gamma}_j'(w_j)\right) - \left(\sum_{j=1}^d g_j(w_j)\right) \right|^{\alpha \wedge 1} + \|\boldsymbol{\gamma} - f\|_\infty \\
\leq &L \left\| \left(\sum_{j=1}^d \left|\boldsymbol{\gamma}_j' - g_j\right|\right)^{\alpha \wedge 1} \right\|_\infty + \|\boldsymbol{\gamma} - f\|_\infty .
\end{aligned} \tag{47}$$

We then apply Lemma 8 and Lemma 9 with $k = \infty$, μ being the Lebesgue measure (probability) and $u(z) = z^{\alpha \wedge 1}$ to (47) and obtain

$$\begin{aligned}
&\left\| \boldsymbol{\gamma}\left(\sum_{j=1}^d \boldsymbol{\gamma}_j'\right) - \left[f\left[(g \vee 0) \wedge 1\right] \vee v_-\right] \wedge v_+ \right\|_\infty \\
\leq &L \sum_{j=1}^d \left\| |\boldsymbol{\gamma}_j' - g_j|^{\alpha \wedge 1} \right\|_\infty + \|\boldsymbol{\gamma} - f\|_\infty \\
\leq &L \sum_{j=1}^d \left(\left\| \boldsymbol{\gamma}_j' - g_j \right\|_\infty \right)^{\alpha \wedge 1} + \|\boldsymbol{\gamma} - f\|_\infty .
\end{aligned} \tag{48}$$

We take

$$r = r(\alpha, \boldsymbol{\beta}) = \left\lfloor \alpha \vee \max_{j=1,\ldots,d} \beta_j \right\rfloor \in \mathbb{N}.$$

By Proposition 6, 7 and Lemma 1, 3, for all $\alpha, L \in \mathbb{R}_+^*$, $\boldsymbol{\beta}, \mathbf{p}, \mathbf{R} \in (\mathbb{R}_+^*)^d$ such that $\beta_j > 1/p_j$, all $(\boldsymbol{l}, t) = (l_1, \ldots, l_d, t) \in \mathbb{N}^d \times \mathbb{N}^*$ and any $\boldsymbol{\gamma}\left(\sum_{j=1}^d \boldsymbol{\gamma}_j'\right) \in \mathcal{F}_{[v_-,v_+]}(\alpha, \boldsymbol{\beta}, \mathbf{p}, L, \mathbf{R})$, we have

$$\inf_{f \in \boldsymbol{S}_{(t,r)}^{\mathcal{H},1}} \|\boldsymbol{\gamma} - f\|_\infty = \inf_{f \in \overline{\boldsymbol{S}}_{(t,r)}^{\mathcal{H},1}} \|\boldsymbol{\gamma} - f\|_\infty \leq C_{\alpha,\boldsymbol{\beta}} L t^{-\alpha} \tag{49}$$

and

$$\inf_{g_j \in \boldsymbol{S}_{(\pi(l_j),r)}^{\mathcal{B},1}} \|\boldsymbol{\gamma}_j' - g_j\|_\infty = \inf_{g_j \in \overline{\boldsymbol{S}}_{(\pi(l_j),r)}^{\mathcal{B},1}} \|\boldsymbol{\gamma}_j' - g_j\|_\infty \leq C_{\alpha,\boldsymbol{\beta},p_j} R_j 2^{-l_j\beta_j}. \tag{50}$$

Combining (48), (49) and (50), we have for all $\alpha, L \in \mathbb{R}_+^*$, $\boldsymbol{\beta}, \mathbf{p}, \mathbf{R} \in (\mathbb{R}_+^*)^d$ such that $\beta_j > 1/p_j$, all $(\boldsymbol{l}, t) = (l_1, \ldots, l_d, t) \in \mathbb{N}^d \times \mathbb{N}^*$ and any $\boldsymbol{\gamma}\left(\sum_{j=1}^d \boldsymbol{\gamma}_j'\right) \in \mathcal{F}_{[v_-,v_+]}(\alpha, \boldsymbol{\beta}, \mathbf{p}, L, \mathbf{R})$,

$$\begin{aligned}
&\inf_{f \in \boldsymbol{S}_{(t,r)}^{\mathcal{H},1},\, g_j \in \boldsymbol{S}_{(\pi(l_j),r)}^{\mathcal{B},1}} \left\| \boldsymbol{\gamma}\left(\sum_{j=1}^d \boldsymbol{\gamma}_j'\right) - \left[f\left[\left(\left(\sum_{j=1}^d g_j\right) \vee 0\right) \wedge 1\right] \vee v_- \right] \wedge v_+ \right\|_{2,P_W}^2 \\
\leq &\inf_{f \in \boldsymbol{S}_{(t,r)}^{\mathcal{H},1},\, g_j \in \boldsymbol{S}_{(\pi(l_j),r)}^{\mathcal{B},1}} \left\| \boldsymbol{\gamma}\left(\sum_{j=1}^d \boldsymbol{\gamma}_j'\right) - \left[f\left[\left(\left(\sum_{j=1}^d g_j\right) \vee 0\right) \wedge 1\right] \vee v_- \right] \wedge v_+ \right\|_{\infty}^2 \\
\leq &C_d \left\{ L^2 \sum_{j=1}^d \left[\left(\inf_{g_j \in \boldsymbol{S}_{(\pi(l_j),r)}^{\mathcal{B},1}} \|\boldsymbol{\gamma}_j' - g_j'\|_\infty \right)^{\alpha \wedge 1} \right]^2 + \inf_{f \in \boldsymbol{S}_{(t,r)}^{\mathcal{H},1}} \|\boldsymbol{\gamma} - f\|_\infty^2 \right\} \\
\leq &C_{d,\alpha,\boldsymbol{\beta},\mathbf{p}} L^2 \left[\sum_{j=1}^d R_j^{2(\alpha\wedge 1)} 2^{-2(\alpha\wedge 1) l_j \beta_j} + t^{-2\alpha} \right].
\end{aligned} \tag{51}$$

We denote $\boldsymbol{\pi}(\boldsymbol{l}) = (\pi(l_1), \ldots, \pi(l_d))$. For any $(l_1, \ldots, l_d, t, r) \in \mathbb{N}^d \times \mathbb{N}^* \times \mathbb{N}$, by Proposition 1 and 6, we have

$$\begin{aligned}
V_{(\boldsymbol{\pi}(\boldsymbol{l}),t,r)}^A + \Delta(\boldsymbol{\pi}(\boldsymbol{l}), t, r) \leq &C \left[\left(t + \sum_{j=1}^d |\pi(l_j)| \right) (r+1) \right] \log(t+r+2) \\
&+ \left[(3\log 2) \left(\sum_{j=1}^d |\pi(l_j)| \right) + r + t \right] \\
\leq &C_r \left(t + \sum_{j=1}^d |\pi(l_j)| \right) \log(t+r+2) \\
\leq &C_{\alpha,\boldsymbol{\beta},\mathbf{p}} \left(t + \sum_{j=1}^d 2^{l_j} \right) \log(t+r+2),
\end{aligned} \tag{52}$$

where C is a numerical constant, C_r is a numerical constant depending only on r and $C_{\alpha,\boldsymbol{\beta},\mathbf{p}}$ is a numerical constant depending only on α, $\boldsymbol{\beta}$, $\mathbf{p}$.

Under Assumption 2, applying (13) together with (51) and (52), we derive that for all $\alpha, L \in \mathbb{R}_+^*$, $\boldsymbol{\beta}, \mathbf{p}, \mathbf{R} \in (\mathbb{R}_+^*)^d$ such that $\beta_j > 1/p_j$, all $(\boldsymbol{l}, t) = (l_1, \ldots, l_d, t) \in \mathbb{N}^d \times \mathbb{N}^*$ and any $\boldsymbol{\gamma}\left(\sum_{j=1}^d \boldsymbol{\gamma}_j'\right) \in \mathcal{F}_{[v_-,v_+]}(\alpha, \boldsymbol{\beta}, \mathbf{p}, L, \mathbf{R})$,

$$\begin{aligned}
\mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}})\right] \leq & C_\kappa \inf_{(\boldsymbol{\pi},t,r)\in (M^{\mathcal{B},1})^d \times (\mathbb{N}^*)^2} \left[\inf_{\widetilde{\boldsymbol{\gamma}} \in \boldsymbol{\Gamma}^A_{(\boldsymbol{\pi},t,r)}} \left\| \boldsymbol{\gamma}\left(\sum_{j=1}^d \boldsymbol{\gamma}_j'\right) - \widetilde{\boldsymbol{\gamma}} \right\|_{2,P_W}^2 \right. \\
& \left. + \frac{\Delta(\boldsymbol{\pi}, t, r)}{n} + \frac{V_{(\boldsymbol{\pi},t,r)}}{n}(1 + \log n) \right] \\
\leq & C_{\kappa,d,\alpha,\boldsymbol{\beta},\mathbf{p}}(1 + \log n) \inf_{(l_1,\ldots,l_d,t)\in \mathbb{N}^d \times \mathbb{N}^*} \left[\left(L^2 t^{-2\alpha} + \frac{t}{n} \right) \right. \\
& \left. + \sum_{j=1}^d \left(L^2 R_j^{2(\alpha\wedge 1)} 2^{-2(\alpha\wedge 1) l_j \beta_j} + \frac{2^{l_j}}{n} \right) \right] \log(t + r + 2). \qquad (53)
\end{aligned}$$

To conclude, we need to optimize the right hand side of (53). We choose $t \geq 1$ such that

$$t - 1 < \left(nL^2\right)^{\frac{1}{1+2\alpha}} \leq t,$$

therefore $L^2 t^{-2\alpha} \leq t/n$ and $t < 1 + \left(nL^2\right)^{\frac{1}{1+2\alpha}}$. As a consequence, we have

$$L^2 t^{-2\alpha} + \frac{t}{n} \leq 2\frac{t}{n} \leq \frac{2}{n} + 2L^{\frac{2}{2\alpha+1}} n^{-\frac{2\alpha}{2\alpha+1}}. \qquad (54)$$

Moreover, we note that if $nL^2 < 1$, we choose $t = 1$, then

$$\log(t + r + 2) \leq \log(r + 3) = C_{\alpha,\boldsymbol{\beta}}. \qquad (55)$$

Otherwise $nL^2 \geq 1$,

$$\begin{aligned}
\log(t + r + 2) & \leq \log\left[\left(nL^2\right)^{\frac{1}{2\alpha+1}} + r + 3\right] \\
& \leq \log\left[C_{\alpha,\boldsymbol{\beta}}\left(nL^2\right)^{\frac{1}{2\alpha+1}}\right] \\
& \leq C_{\alpha,\boldsymbol{\beta}}\left(\log n \vee \log L^2 \vee 1\right). \qquad (56)
\end{aligned}$$

For any $j \in \{1, \ldots, d\}$, if $nL^2 R_j^{2(\alpha\wedge 1)} < 1$, we take $l_j = 0$ so that

$$L^2 R_j^{2(\alpha\wedge 1)} 2^{-2(\alpha\wedge 1) l_j \beta_j} + \frac{2^{l_j}}{n} < \frac{2}{n}. \qquad (57)$$

Otherwise, we take l_j as the largest natural number such that

$$\frac{2^{l_j}}{n} \leq L^2 R_j^{2(\alpha\wedge 1)} 2^{-2(\alpha\wedge 1) l_j \beta_j},$$

which yields

$$
\begin{aligned}
L^2 R_j^{2(\alpha\wedge 1)} 2^{-2(\alpha\wedge 1)l_j\beta_j} + \frac{2^{l_j}}{n} &\leq L^2 R_j^{2(\alpha\wedge 1)} 2^{1-2(\alpha\wedge 1)l_j\beta_j} \\
&\leq C_{\alpha,\boldsymbol{\beta}} \left[L^2 R_j^{2(\alpha\wedge 1)}\right]^{\frac{1}{2(\alpha\wedge 1)\beta_j+1}} n^{-\frac{2(\alpha\wedge 1)\beta_j}{2(\alpha\wedge 1)\beta_j+1}} \\
&\leq C_{\alpha,\boldsymbol{\beta}} \left(L R_j^{\alpha\wedge 1}\right)^{\frac{2}{2(\alpha\wedge 1)\beta_j+1}} n^{-\frac{2(\alpha\wedge 1)\beta_j}{2(\alpha\wedge 1)\beta_j+1}}. \qquad (58)
\end{aligned}
$$

Combining (53), (54), (55), (56), (57) and (58), we obtain whatever the distribution of W, for all $\alpha, L \in \mathbb{R}_+^*$, $\boldsymbol{\beta}, \mathbf{p}, \mathbf{R} \in (\mathbb{R}_+^*)^d$ such that $\beta_j > 1/p_j$ and any $\boldsymbol{\gamma}^\star \in \mathcal{F}_{[v_-,v_+]}(\alpha, \boldsymbol{\beta}, \mathbf{p}, L, \mathbf{R})$,

$$
\begin{aligned}
&C'_{\kappa,d,\alpha,\boldsymbol{\beta},\mathbf{p}} \mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}})\right] \\
&\leq \left\{ \left[\sum_{j=1}^d \left(L R_j^{\alpha\wedge 1}\right)^{\frac{2}{2(\alpha\wedge 1)\beta_j+1}} n^{-\frac{2(\alpha\wedge 1)\beta_j}{2(\alpha\wedge 1)\beta_j+1}} \right] + L^{\frac{2}{2\alpha+1}} n^{-\frac{2\alpha}{2\alpha+1}} + \frac{1}{n} \right\} \mathcal{L}_n^2,
\end{aligned}
$$

where $\mathcal{L}_n = \log n \vee \log L^2 \vee 1$. Finally, the conclusion follows by taking the supremum over $\mathcal{F}_{[v_-,v_+]}(\alpha, \boldsymbol{\beta}, \mathbf{p}, L, \mathbf{R})$.

A.4 Proof of Corollary 3. We first present the following result which can be proved by a similar argument as the proof of Lemma 1.

Lemma 10. *Let* $\mathcal{C}_d = \left\{(c_1, \ldots, c_d) \in \mathbb{R}^d,\ \sum_{j=1}^d |c_j| \leq 1\right\}$. *We denote* $\overline{\boldsymbol{S}}_{\mathcal{C}_d}$ *the collection of functions on* $[0,1]^d$ *of the form*

$$
f(\boldsymbol{w}) = \frac{1}{2}\left(\langle c, \boldsymbol{w}\rangle + 1\right), \quad \textit{for all} \quad \boldsymbol{w} \in [0,1]^d, \qquad (59)
$$

with $c \in \mathcal{C}_d$ *and* $\boldsymbol{S}_{\mathcal{C}_d}$ *the collection of functions of the form in (59) but with* $c \in \mathcal{C}_d \cap \mathbb{Q}^d$. *Then* $\boldsymbol{S}_{\mathcal{C}_d}$ *is dense in* $\overline{\boldsymbol{S}}_{\mathcal{C}_d}$ *with respect to the supremum norm.*

Now let us turn to prove Corollary 3. For all $\boldsymbol{\alpha} \in (\mathbb{R}_+^*)^l$, $L > 0$, any $\boldsymbol{\gamma}^\star = \boldsymbol{\gamma} \circ \boldsymbol{\gamma}' \in \mathcal{G}_{[v_-,v_+]}(\boldsymbol{\alpha}, L)$, where $\boldsymbol{\gamma}'(\boldsymbol{w}) = (\gamma_1'(\boldsymbol{w}), \ldots, \gamma_l'(\boldsymbol{w}))$ with $\gamma_j' \in \overline{\boldsymbol{S}}_{\mathcal{C}_d}$ for $j \in \{1, \ldots, l\}$, $\boldsymbol{\gamma} \in \mathcal{H}^{\boldsymbol{\alpha}}(L, v_-, v_+)$ and any $f \in \boldsymbol{S}^{\mathcal{H},l}_{(\boldsymbol{t},r)}$, $g : [0,1]^d \to [0,1]^l$ defined as $g(\boldsymbol{w}) = (g_1(\boldsymbol{w}), \ldots, g_l(\boldsymbol{w}))$ with $g_j \in \boldsymbol{S}_{\mathcal{C}_d}$ for $j \in \{1, \ldots, l\}$, we have

$$
\begin{aligned}
\|\boldsymbol{\gamma} \circ \boldsymbol{\gamma}' - (f \circ g) \vee v_-) \wedge v_+\|_\infty &\leq \|\boldsymbol{\gamma} \circ \boldsymbol{\gamma}' - f \circ g\|_\infty \\
&\leq \|\boldsymbol{\gamma} \circ \boldsymbol{\gamma}' - \boldsymbol{\gamma} \circ g\|_\infty + \|\boldsymbol{\gamma} \circ g - f \circ g\|_\infty \\
&\leq \left\| L \sum_{j=1}^l \left|\gamma_j' - g_j\right|^{\alpha_j\wedge 1} \right\|_\infty + \|\boldsymbol{\gamma} - f\|_\infty \\
&\leq L \sum_{j=1}^l \left\| \left|\gamma_j' - g_j\right|^{\alpha_j\wedge 1} \right\|_\infty + \|\boldsymbol{\gamma} - f\|_\infty. \qquad (60)
\end{aligned}
$$

We apply Lemma 9 to (60) by taking $k=\infty$, μ the Lebesgue probability and $u(z)=z^{\alpha_j\wedge 1}$ for each $j\in\{1,\dots,l\}$ and obtain

$$\|\boldsymbol{\gamma}\circ\boldsymbol{\gamma}'-(f\circ g\vee v_-)\wedge v_+\|_\infty \le L\sum_{j=1}^{l}\left(\left\|\boldsymbol{\gamma}_j'-g_j\right\|_\infty\right)^{\alpha_j\wedge 1}+\|\boldsymbol{\gamma}-f\|_\infty. \tag{61}$$

We take $r=\max_{j=1,\dots,l}\lfloor\alpha_j\rfloor\in\mathbb{N}$. By Proposition 7, for any $\boldsymbol{\gamma}^\star=\boldsymbol{\gamma}\circ\boldsymbol{\gamma}'\in\mathcal{G}_{[v_-,v_+]}(\boldsymbol{\alpha},L)$ and all $\boldsymbol{t}=(t_1,\dots,t_l)\in(\mathbb{N}^*)^l$, we have

$$\inf_{f\in\overline{\boldsymbol{S}}^{\mathcal{H},l}_{(\boldsymbol{t},r)}}\|\boldsymbol{\gamma}-f\|_\infty \le C_{l,\boldsymbol{\alpha}}L\sum_{j=1}^{l}t_j^{-\alpha_j},$$

where $C_{l,\boldsymbol{\alpha}}$ is a constant depending on l and $\boldsymbol{\alpha}$ only. Then by Lemma 3, $\boldsymbol{S}^{\mathcal{H},l}_{(\boldsymbol{t},r)}$ is dense in $\overline{\boldsymbol{S}}^{\mathcal{H},l}_{(\boldsymbol{t},r)}$ with respect to the supremum norm $\|\cdot\|_\infty$, we obtain

$$\inf_{f\in\boldsymbol{S}^{\mathcal{H},l}_{(\boldsymbol{t},r)}}\|\boldsymbol{\gamma}-f\|_\infty=\inf_{f\in\overline{\boldsymbol{S}}^{\mathcal{H},l}_{(\boldsymbol{t},r)}}\|\boldsymbol{\gamma}-f\|_\infty\le C_{l,\boldsymbol{\alpha}}L\sum_{j=1}^{l}t_j^{-\alpha_j}. \tag{62}$$

Therefore, following from (61), (62) and Lemma 10, for all $\boldsymbol{\alpha}\in(\mathbb{R}_+^*)^l$, $L>0$, any $\boldsymbol{\gamma}^\star=\boldsymbol{\gamma}\circ\boldsymbol{\gamma}'\in\mathcal{G}_{[v_-,v_+]}(\boldsymbol{\alpha},L)$ and all $\boldsymbol{t}=(t_1,\dots,t_l)\in(\mathbb{N}^*)^l$,

$$\begin{aligned}
&\inf_{f\in\boldsymbol{S}^{\mathcal{H},l}_{(\boldsymbol{t},r)},\, g_j\in\boldsymbol{S}_{\mathcal{C}_d}}\|\boldsymbol{\gamma}\circ\boldsymbol{\gamma}'-(f\circ g\vee v_-)\wedge v_+\|^2_{2,P_W}\\
\le&\inf_{f\in\boldsymbol{S}^{\mathcal{H},l}_{(\boldsymbol{t},r)},\, g_j\in\boldsymbol{S}_{\mathcal{C}_d}}\|\boldsymbol{\gamma}\circ\boldsymbol{\gamma}'-(f\circ g\vee v_-)\wedge v_+\|^2_{\infty}\\
\le&C_{l,\boldsymbol{\alpha}}L^2\left(\sum_{j=1}^{l}t_j^{-\alpha_j}\right)^2. \qquad (63)
\end{aligned}$$

Moreover, for any $r\in\mathbb{N}$ and $\boldsymbol{t}=(t_1,\dots,t_l)\in(\mathbb{N}^*)^l$, with the fact that

$$\Delta(\boldsymbol{t},r)=\sum_{j=1}^{l}t_j+r\le l\prod_{j=1}^{l}t_j+r\le l\left(\prod_{j=1}^{l}t_j\right)(r+1)^l$$

and Proposition 2, we have

$$\begin{aligned}
V^M_{(\boldsymbol{t},r)}+\Delta(\boldsymbol{t},r)&\le C_l\left[d+\left(\prod_{j=1}^{l}t_j\right)(r+1)^l\right]\log\left[\left(\sum_{j=1}^{l}t_j\right)+lr+l+1\right]\\
&\le C_{l,\boldsymbol{\alpha}}\left[d+\left(\prod_{j=1}^{l}t_j\right)\right]\log\left[\left(\sum_{j=1}^{l}t_j\right)+lr+l+1\right], \qquad (64)
\end{aligned}$$

where C_l is a numerical constant depending only on l and $C_{l,\boldsymbol{\alpha}}$ is a numerical constant depending on l, $\boldsymbol{\alpha}$ only

Under Assumption 2, applying (13) together with the inequalities (63) and (64), we derive that for all $\boldsymbol{\alpha}\in(\mathbb{R}_+^*)^l$ and $L>0$, any $\boldsymbol{\gamma}^\star=\boldsymbol{\gamma}\circ\boldsymbol{\gamma}'\in$

$\mathcal{G}_{[v_-,v_+]}(\boldsymbol{\alpha}, L)$, whatever the distribution of W,

$$\begin{aligned}
&\mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}})\right] \\
&\leq C_{\boldsymbol{\kappa}} \inf_{(\boldsymbol{t},r)\in(\mathbb{N}^*)^l\times\mathbb{N}} \left[\inf_{\widetilde{\boldsymbol{\gamma}}\in\boldsymbol{\Gamma}^M_{(\boldsymbol{t},r)}} \|\boldsymbol{\gamma}\circ\boldsymbol{\gamma}' - \widetilde{\boldsymbol{\gamma}}\|^2_{2,P_W} + \frac{\Delta(\boldsymbol{t},r)}{n} + \frac{V^M_{(\boldsymbol{t},r)}}{n}(1+\log n)\right] \\
&\leq C_{\boldsymbol{\kappa},l,\boldsymbol{\alpha}}(1+\log n) \inf_{\boldsymbol{t}\in(\mathbb{N}^*)^l} \left[L^2\left(\sum_{j=1}^l t_j^{-\alpha_j}\right)^2 + \frac{\prod_{j=1}^l t_j}{n} + \frac{d}{n}\right]\log(U),
\end{aligned} \tag{65}$$

where $U = \sum_{j=1}^l t_j + lr + l + 1$. We then optimize the risk bound given on the right hand side of (65). For each $j \in \{1,\dots,l\}$, we choose $t_j \geq 1$ satisfying

$$t_j - 1 < (nL^2)^{\frac{\overline{\alpha}}{(2\overline{\alpha}+l)\alpha_j}} \leq t_j,$$

where $\overline{\alpha}$ denotes the harmonic mean of $\alpha_1,\dots,\alpha_l$. Therefore, we have

$$L^2\left(\sum_{j=1}^l t_j^{-\alpha_j}\right)^2 \leq l^2L^2(nL^2)^{-\frac{2\overline{\alpha}}{2\overline{\alpha}+l}} = l^2L^{\frac{2l}{2\overline{\alpha}+l}}n^{-\frac{2\overline{\alpha}}{2\overline{\alpha}+l}}. \tag{66}$$

If $nL^2 \leq 1$, then $t_j = 1$ for all $j \in \{1,\dots,l\}$ hence

$$\frac{\prod_{j=1}^l t_j}{n} \leq \frac{1}{n} \tag{67}$$

and for some numerical constant $C_{l,\boldsymbol{\alpha}}$ depending on l and $\boldsymbol{\alpha}$ only

$$\log(U) = \log\left[\left(\sum_{j=1}^l t_j\right) + lr + l + 1\right] \leq C_{l,\boldsymbol{\alpha}}. \tag{68}$$

Otherwise,

$$\begin{aligned}
\frac{\prod_{j=1}^l t_j}{n} &\leq \frac{\prod_{j=1}^l 2(nL^2)^{\frac{\overline{\alpha}}{(2\overline{\alpha}+l)\alpha_j}}}{n} = \frac{2^l(nL^2)^{\frac{\overline{\alpha}}{2\overline{\alpha}+l}\sum_{j=1}^l\frac{1}{\alpha_j}}}{n} \\
&\leq 2^l\frac{(nL^2)^{\frac{l}{2\overline{\alpha}+l}}}{n} \leq 2^lL^{\frac{2l}{2\overline{\alpha}+l}}n^{-\frac{2\overline{\alpha}}{2\overline{\alpha}+l}}
\end{aligned} \tag{69}$$

and

$$\begin{aligned}
\log\left[\left(\sum_{j=1}^l t_j\right) + lr + l + 1\right] &\leq \log\left[l\left(\prod_{j=1}^l t_j\right) + lr + l + 1\right] \\
&\leq \log\left[C_l\left(nL^2\right)^{\frac{l}{2\overline{\alpha}+l}} + C_{l,\boldsymbol{\alpha}}\right] \\
&\leq C_{l,\boldsymbol{\alpha}}\left(\log n \vee \log L^2 \vee 1\right).
\end{aligned} \tag{70}$$

Plugging (66), (67), (68), (69), (70) into (65), we have that whatever the distribution of W, for all $\boldsymbol{\alpha} \in (\mathbb{R}_+^*)^l$ and $L > 0$, any $\boldsymbol{\gamma}^\star \in \mathcal{G}_{[v_-,v_+]}(\boldsymbol{\alpha}, L)$

$$\mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}})\right] \leq C_{\kappa,l,\boldsymbol{\alpha}} \left(L^{\frac{2l}{2\overline{\alpha}+l}} n^{-\frac{2\overline{\alpha}}{2\overline{\alpha}+l}} + \frac{d}{n} \right) \left(\log n \vee \log L^2 \vee 1\right)^2, \tag{71}$$

where $C_{\kappa,l,\boldsymbol{\alpha}}$ is a constant depending only on κ, l and $\boldsymbol{\alpha}$. The conclusion finally follows by taking the supremum over the set $\mathcal{G}_{[v_-,v_+]}(\boldsymbol{\alpha}, L)$.

A.5 Proof of Corollary 4. For any $\boldsymbol{\gamma}^\star \in \mathcal{F}_{[v_-,v_+]}(t, l, \mathbf{p}, K)$, we first rewrite it as $\boldsymbol{\gamma}^\star = \sum_{k\in\mathbb{N}^*} t^k \boldsymbol{\gamma}_1 \left(\boldsymbol{\gamma}_2^{\circ k}\right)$, where $t \in (-1,1)$, $\boldsymbol{\gamma}_1 \in \boldsymbol{S}_{(l,p_1)}$ and $\boldsymbol{\gamma}_2 \in \boldsymbol{S}_{(l,p_2)}$. For any $m \in \mathbb{N}^*$, we denote $\boldsymbol{\gamma}_m^\star = \sum_{k=1}^m t^k \boldsymbol{\gamma}_1 \left(\boldsymbol{\gamma}_2^{\circ k}\right)$ the m-partial sum of the function $\boldsymbol{\gamma}^\star$. We then apply Proposition 4.4 of Daubechies et al. (2019) and obtain that $\boldsymbol{\gamma}_m^\star \in \overline{\boldsymbol{S}}_{(l(m+1),p_1+p_2+2)}$. We note that for any $\boldsymbol{\gamma}_m^\star = \sum_{k=1}^m t^k \boldsymbol{\gamma}_1 \left(\boldsymbol{\gamma}_2^{\circ k}\right)$, there exists a sequence of functions $\{\boldsymbol{\gamma}_i\}_{i\in\mathbb{N}}$ with $\boldsymbol{\gamma}_i = \sum_{k=1}^m t_i^k \boldsymbol{\gamma}_1 \left(\boldsymbol{\gamma}_2^{\circ k}\right) \in \boldsymbol{S}_{(l(m+1),p_1+p_2+2)}$ and $t_i \in (-1,1) \cap \mathbb{Q}$ such that

$$\begin{aligned}
\lim_{i\to+\infty} \|\boldsymbol{\gamma}_m^\star - \boldsymbol{\gamma}_i\|_\infty &= \lim_{i\to+\infty} \left| \sum_{k=1}^m \left(t^k - t_i^k\right) \boldsymbol{\gamma}_1 \left(\boldsymbol{\gamma}_2^{\circ k}\right)(w) \right| \\
&\leq K \lim_{i\to+\infty} \sum_{k=1}^m \left|t^k - t_i^k\right| \\
&\leq \frac{Km(m+1)}{2} \lim_{i\to+\infty} |t - t_i| = 0, \qquad (72)
\end{aligned}$$

since $\mathbb{Q}$ is dense in $\mathbb{R}$. Therefore, with the fact that $\boldsymbol{\gamma}^\star$ taking values in $[v_-, v_+]$ and (72), we have

$$\begin{aligned}
\inf_{\overline{\boldsymbol{\gamma}}\in\boldsymbol{\Gamma}_{(l(m+1),p_1+p_2+2)}} \|\boldsymbol{\gamma}^\star - \overline{\boldsymbol{\gamma}}\|_\infty &\leq \inf_{\overline{\boldsymbol{\gamma}}\in\boldsymbol{S}_{(l(m+1),p_1+p_2+2)}} \|\boldsymbol{\gamma}^\star - \overline{\boldsymbol{\gamma}}\|_\infty \\
&\leq \|\boldsymbol{\gamma}^\star - \boldsymbol{\gamma}_m^\star\|_\infty + \inf_{\overline{\boldsymbol{\gamma}}\in\boldsymbol{S}_{(l(m+1),p_1+p_2+2)}} \|\boldsymbol{\gamma}_m^\star - \overline{\boldsymbol{\gamma}}\|_\infty \\
&\leq \sup_{w\in[0,1]} \left| \sum_{k=m+1}^{+\infty} t^k \boldsymbol{\gamma}_1 \left(\boldsymbol{\gamma}_2^{\circ k}(w)\right) \right| \\
&\leq C_{t,K} |t|^{m+1}, \qquad (73)
\end{aligned}$$

where $C_{t,K}$ stands for a numerical constant depending on t and K only. We denote $V_{(l(m+1),p_1+p_2+2)}$ the VC dimension of $\boldsymbol{\Gamma}_{(l(m+1),p_1+p_2+2)}$. With the fact that $\boldsymbol{S}_{(l(m+1),p_1+p_2+2)} \subset \overline{\boldsymbol{S}}_{(l(m+1),p_1+p_2+2)}$, Proposition 3 and

$$\boldsymbol{\Gamma}_{(l(m+1),p_1+p_2+2)} = \left\{(\boldsymbol{\gamma} \vee v_-) \wedge v_+,\ \boldsymbol{\gamma} \in \boldsymbol{S}_{(l(m+1),p_1+p_2+2)}\right\},$$

we derive that for some numerical constant C

$$V_{(l(m+1),p_1+p_2+2)} \leq C l_0 \left[p_0^2(l_0-1) + p_0(l_0+2) + 1\right] \log\left[(l_0+1)\left(\frac{p_0 l_0}{2} + 1\right)\right],$$

where $l_0 = l(m+1)$ and $p_0 = p_1 + p_2 + 2$. Then it follows by a basic computation that

$$V_{(l(m+1),p_1+p_2+2)} + \Delta(l(m+1), p_1+p_2+2) \leq C_{l,\mathbf{p}}(m+1)^3, \tag{74}$$

where $C_{l,\mathbf{p}}$ is a numerical constant depending on l and $\mathbf{p}$ only.

We take $L = l(m+1)$ and $p = p_1+p_2+2$. Under Assumption 2, applying (13) together with the inequalities (73) and (74), we have no matter what the distribution of W is, for all $t \in (-1,1)$, $l \in \mathbb{N}^*$, $\mathbf{p} \in (\mathbb{N}^*)^2$ and $K \geq 0$, for any $\boldsymbol{\gamma}^\star \in \mathcal{F}_{[v_-,v_+]}(t,l,\mathbf{p},K)$,

$$\begin{aligned}\mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}})\right] \leq & C_{\kappa,\mathbf{p},l} \inf_{m\in\mathbb{N}^*}\left[\inf_{\overline{\boldsymbol{\gamma}}\in\boldsymbol{\Gamma}_{(l(m+1),p_1+p_2+2)}} \|\boldsymbol{\gamma}^\star - \overline{\boldsymbol{\gamma}}\|^2_{2,P_W} \right.\\ & \left. + \frac{(m+1)^3}{n}(1+\log n)\right]\\ \leq & C_{\kappa,\mathbf{p},l,t,K} \inf_{m\in\mathbb{N}^*}\left[|t|^{2(m+1)} + \frac{(m+1)^3}{n}(1+\log n)\right]. \end{aligned} \tag{75}$$

Now we only need to optimize the right hand side of (75). If $|t|^4 \leq 1/n$, we choose $m = 1$ so that

$$\mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}})\right] \leq C_{\kappa,\mathbf{p},l,t,K}\frac{1}{n}(1+\log n). \tag{76}$$

Otherwise, we choose $m \in \mathbb{N}^*$ such that

$$m < \frac{\log n}{-2\log|t|} \leq m+1.$$

With this choice, $|t|^{2(m+1)} \leq 1/n$ and we derive from (75),

$$\mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}})\right] \leq C_{\kappa,\mathbf{p},l,t,K}\frac{(1+\log n)^4}{n}. \tag{77}$$

Combining the results in (76) and (77), whatever the distribution of W, for all $t \in (-1,1)$, $l \in \mathbb{N}^*$, $\mathbf{p} \in (\mathbb{N}^*)^2$ and $K \geq 0$, any $\boldsymbol{\gamma}^\star \in \mathcal{F}_{[v_-,v_+]}(t,l,\mathbf{p},K)$, we have

$$\mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}})\right] \leq C_{\kappa,\mathbf{p},l,t,K}\frac{1}{n}(1+\log n)^4. \tag{78}$$

Then the conclusion follows by taking the supremum over $\mathcal{F}_{[v_-,v_+]}(t,l,\mathbf{p},K)$ on both sides of (78).

A.6 Proof of Corollary 5. For any $(L,p,\boldsymbol{s}) \in (\mathbb{N}^*)^2 \times \{0,1\}^{\overline{p}}$, let $\overline{\boldsymbol{S}}'_{(L,p,\boldsymbol{s})} \subset \overline{\boldsymbol{S}}_{(L,p,\boldsymbol{s})}$ be the collection of functions based on sparse neural network, where all the non-zero parameters vary in $[-1,1]$ and $\boldsymbol{S}'_{(L,p,\boldsymbol{s})} \subset \overline{\boldsymbol{S}}'_{(L,p,\boldsymbol{s})}$ be the collection of functions, where all the non-zero parameters vary in $[-1,1]\cap\mathbb{Q}$. We first show the following result.

Lemma 11. *For any $(L,p,\boldsymbol{s}) \in (\mathbb{N}^*)^2 \times \{0,1\}^{\overline{p}}$, $\boldsymbol{S}'_{(L,p,\boldsymbol{s})}$ is dense in $\overline{\boldsymbol{S}}'_{(L,p,\boldsymbol{s})}$ with respect to the supremum norm $\|\cdot\|_\infty$.*

Proof. By the definition of dense with respect to the supremum norm, we need to show that for any function $f \in \overline{\boldsymbol{S}}'_{(L,p,\boldsymbol{s})}$, there is a sequence of functions $f_i \in \boldsymbol{S}'_{(L,p,\boldsymbol{s})}$, $i \in \mathbb{N}$ such that

$$\lim_{i \to +\infty} \|f - f_i\|_\infty = 0.$$

The idea is inspired by the proof of Lemma 5 of Schmidt-Hieber (2020). Recall for any $f \in \overline{\boldsymbol{S}}_{(L,p)}$, it can be written as

$$f(\boldsymbol{w}) = M_L \circ \sigma \circ M_{L-1} \circ \cdots \circ \sigma \circ M_0(\boldsymbol{w}), \quad \text{for all } \boldsymbol{w} \in [0,1]^d.$$

For $l \in \{0, \ldots, L+1\}$, we define $p_l = p$ for $l \in \{1, \ldots, L\}$, $p_0 = d$ and $p_{L+1} = 1$. For $l \in \{1, \ldots, L\}$, we define the function $f_l^+ : [0,1]^d \to \mathbb{R}^p$,

$$f_l^+(\boldsymbol{w}) = \sigma \circ M_{l-1} \circ \cdots \circ \sigma \circ M_0(\boldsymbol{w})$$

and for $l \in \{1, \ldots, L+1\}$, we define $f_l^- : \mathbb{R}^{p_{l-1}} \to \mathbb{R}$

$$f_l^-(\boldsymbol{x}) = M_L \circ \sigma \circ \cdots \circ \sigma \circ M_{l-1}(\boldsymbol{x}).$$

We set the notations $f_0^+(\boldsymbol{w}) = f_{L+2}^-(\boldsymbol{w}) = \boldsymbol{w}$. Given a vector $\boldsymbol{v} = (v_1, \ldots, v_p)$ of any size $p \in \mathbb{N}^*$, we denote $|\boldsymbol{v}|_\infty = \max_{i=1,\ldots,p} |v_i|$.

For any $f \in \overline{\boldsymbol{S}}'_{(L,p,\boldsymbol{s})}$, with the fact that the absolute values of all the parameters are bounded by 1 and $\boldsymbol{w} \in [0,1]^d$, we have for all $l \in \{1, \ldots, L\}$

$$\left|f_l^+(\boldsymbol{w})\right|_\infty \leq \prod_{k=0}^{l-1} (p_k + 1)$$

and f_l^-, $l \in \{1, \ldots, L+1\}$, is a multivariate Lipschitz function with Lipschitz constant bounded by $\prod_{k=l-1}^{L} p_k$.

For any $f \in \overline{\boldsymbol{S}}'_{(L,p,\boldsymbol{s})}$ with weight matrices and shift vectors $\{M_l = (A_l, b_l)\}_{l=0}^L$ and for all $\epsilon > 0$, since $\mathbb{Q}$ is dense in $\mathbb{R}$, there exist a $N_\epsilon > 0$ such that for all $i \geq N_\epsilon$, all the non-zero parameters in $f_i \in \boldsymbol{S}'_{(L,p,\boldsymbol{s})}$ are smaller than $\epsilon/(L+1)\left[\prod_{k=0}^{L+1}(p_k+1)\right]$ away from the corresponding ones in f. We denote the weight matrices and shift vectors of function f_i as $\{M_l^i = (A_l^i, b_l^i)\}_{l=0}^L$. We note that

$$f_i(\boldsymbol{w}) = f_{i,2}^- \circ \sigma \circ M_0^i \circ f_0^+(\boldsymbol{w})$$

and

$$f(\boldsymbol{w}) = f_{i,L+2}^- \circ M_L \circ f_L^+(\boldsymbol{w}).$$

Therefore, for all $i \geq N_\epsilon$ and all $\boldsymbol{w} \in [0,1]^d$

$$
\begin{aligned}
|f_i(\boldsymbol{w}) - f(\boldsymbol{w})| \leq & \sum_{l=1}^{L} \left| f^-_{i,l+1} \circ \sigma \circ M^i_{l-1} \circ f^+_{l-1}(\boldsymbol{w}) - f^-_{i,l+1} \circ \sigma \circ M_{l-1} \circ f^+_{l-1}(\boldsymbol{w}) \right| \\
& + \left| M^i_L \circ f^+_L(\boldsymbol{w}) - M_L \circ f^+_L(\boldsymbol{w}) \right| \\
\leq & \sum_{l=1}^{L} \left(\prod_{k=l}^{L} p_k \right) \left| M^i_{l-1} \circ f^+_{l-1}(\boldsymbol{w}) - M_{l-1} \circ f^+_{l-1}(\boldsymbol{w}) \right|_\infty \\
& + \left| M^i_L \circ f^+_L(\boldsymbol{w}) - M_L \circ f^+_L(\boldsymbol{w}) \right| \\
\leq & \sum_{l=1}^{L+1} \left(\prod_{k=l}^{L+1} p_k \right) \left| M^i_{l-1} \circ f^+_{l-1}(\boldsymbol{w}) - M_{l-1} \circ f^+_{l-1}(\boldsymbol{w}) \right|_\infty \\
\leq & \sum_{l=1}^{L+1} \left(\prod_{k=l}^{L+1} p_k \right) \left[\left| \left(A^i_{l-1} - A_{l-1} \right) \circ f^+_{l-1}(\boldsymbol{w}) \right|_\infty + |b^i_{l-1} - b_{l-1}|_\infty \right] \\
< & \frac{\epsilon}{(L+1)\left[\prod_{k=0}^{L+1}(p_k+1)\right]} \sum_{l=1}^{L+1} \left(\prod_{k=l}^{L+1} p_k \right) \left(p_{l-1} \left| f^+_{l-1}(\boldsymbol{w}) \right|_\infty + 1 \right) \\
< & \epsilon.
\end{aligned}
$$

Hence, by the definition we can conclude that $\boldsymbol{S}'_{(L,p,\boldsymbol{s})}$ is dense in $\overline{\boldsymbol{S}}'_{(L,p,\boldsymbol{s})}$ with respect to the supremum norm $\|\cdot\|_\infty$. □

Then, we borrow the approximation result, more precisely (25) and (26), in the proof of Theorem 1 of Schmidt-Hieber (2020). For all $k \in \mathbb{N}^*$, $K \geq 0$, $\mathbf{d} \in (\mathbb{N}^*)^{k+1}$, $\mathbf{t} \in (\mathbb{N}^*)^{k+1}$ with $t_j \leq d_j$ for $j \in \{0, \ldots, k\}$, $\boldsymbol{\alpha} \in (\mathbb{R}^*_+)^{k+1}$ and all $\boldsymbol{\gamma}^\star \in \mathcal{F}_{[v_-,v_+]}(k, \mathbf{d}, \mathbf{t}, \boldsymbol{\alpha}, K)$, there exists a sparse neural network which can be embedded into $\overline{\boldsymbol{S}}'_{(L,p,\boldsymbol{s})}$, for sufficiently large n, satisfying

(i) $\sum_{i=0}^{k} \log_2(4t_i + 4\alpha_i) \log_2 n \leq L \lesssim n\phi_n$,
(ii) $n\phi_n \lesssim p$,
(iii) $\|\boldsymbol{s}\|_0 \asymp n\phi_n \log n$,

such that

$$
\inf_{\overline{\boldsymbol{\gamma}} \in \overline{\boldsymbol{S}}'_{(L,p,\boldsymbol{s})}} \|\boldsymbol{\gamma}^\star - \overline{\boldsymbol{\gamma}}\|^2_\infty \leq C_{k,\mathbf{d},\mathbf{t},\boldsymbol{\alpha},K} \max_{i=0,\ldots,k} n^{-\frac{2\alpha'_i}{2\alpha'_i + t_i}},
$$

where $C_{k,\mathbf{d},\mathbf{t},\boldsymbol{\alpha},K}$ is a numerical constant depending only on k, $\mathbf{d}$, $\mathbf{t}$, $\boldsymbol{\alpha}$ and K. Moreover, with the fact that $\boldsymbol{\gamma}^\star$ taking values in $[v_-, v_+]$ and Lemma 11,

we have

$$\begin{aligned}\inf_{\overline{\boldsymbol{\gamma}}\in\overline{\boldsymbol{\Gamma}}_{(L,p,\boldsymbol{s})}}\|\boldsymbol{\gamma}^\star-\overline{\boldsymbol{\gamma}}\|_\infty^2 &\le \inf_{\overline{\boldsymbol{\gamma}}\in\boldsymbol{S}_{(L,p,\boldsymbol{s})}}\|\boldsymbol{\gamma}^\star-\overline{\boldsymbol{\gamma}}\|_\infty^2 \le \inf_{\overline{\boldsymbol{\gamma}}\in\boldsymbol{S}'_{(L,p,\boldsymbol{s})}}\|\boldsymbol{\gamma}^\star-\overline{\boldsymbol{\gamma}}\|_\infty^2\\ &\le \inf_{\overline{\boldsymbol{\gamma}}\in\overline{\boldsymbol{S}}'_{(L,p,\boldsymbol{s})}}\|\boldsymbol{\gamma}^\star-\overline{\boldsymbol{\gamma}}\|_\infty^2\\ &\le C_{k,\mathbf{d},\mathbf{t},\boldsymbol{\alpha},K}\max_{i=0,\dots,k} n^{-\frac{2\alpha'_i}{2\alpha'_i+t_i}}. \qquad (79)\end{aligned}$$

Let $C_{k,\mathbf{t},\boldsymbol{\alpha}}$ and $C'_{k,\mathbf{d},\mathbf{t},\boldsymbol{\alpha}}$ be two numerical constants depending only on their subscripts. We choose

$$L = C_{k,\mathbf{t},\boldsymbol{\alpha}}\log_2 n \quad\text{and}\quad p = C'_{k,\mathbf{d},\mathbf{t},\boldsymbol{\alpha}} n\phi_n,$$

which satisfy the conditions (i) and (ii) for n large enough. It follows by Proposition 3 and the definition of $\overline{\boldsymbol{\Gamma}}_{(L,p,\boldsymbol{s})}$ that for n sufficiently large, the VC dimension $V_{(L,p,\boldsymbol{s})}$ of $\overline{\boldsymbol{\Gamma}}_{(L,p,\boldsymbol{s})}$ satisfies

$$\begin{aligned}V_{(L,p,\boldsymbol{s})} &\le C_{k,\mathbf{d},\mathbf{t},\boldsymbol{\alpha}} n\phi_n(\log n)^2\log\left[(L+1)\left(\frac{pL}{2}+1\right)\right]\\ &\le C_{k,\boldsymbol{d},\boldsymbol{t},\boldsymbol{\alpha}} n\phi_n\left(\log n\right)^3. \qquad (80)\end{aligned}$$

Moreover, with our choices of $L, p, \boldsymbol{s}$ and n large enough,

$$\begin{aligned}\Delta(L,p,\boldsymbol{s}) &\le \|\boldsymbol{s}\|_0\log\left(2e\overline{p}\right)+p+L\\ &\le C_{k,\boldsymbol{d},\boldsymbol{t},\boldsymbol{\alpha}}\left(n\phi_n\log n\log\overline{p}+n\phi_n+\log_2 n\right)\\ &\le C_{k,\boldsymbol{d},\boldsymbol{t},\boldsymbol{\alpha}} n\phi_n(\log n)^2. \qquad (81)\end{aligned}$$

Under Assumption 2, applying (13) together with (79), (80) and (81), whatever the distribution of W, we derive that for all $k\in\mathbb{N}^*$, $K\ge 0$, $\mathbf{d}\in(\mathbb{N}^*)^{k+1}$, $\mathbf{t}\in(\mathbb{N}^*)^{k+1}$ with $t_j\le d_j$ for $j\in\{0,\dots,k\}$ and $\boldsymbol{\alpha}\in(\mathbb{R}_+^*)^{k+1}$, any $\boldsymbol{\gamma}^\star\in\mathcal{F}_{[v_-,v_+]}(k,\mathbf{d},\mathbf{t},\boldsymbol{\alpha},K)$, with a sufficiently large n

$$\begin{aligned}\mathbb{E}\left[h^2\left(R_{\boldsymbol{\gamma}^\star},R_{\widehat{\boldsymbol{\gamma}}}\right)\right] &\le C_\kappa\left[\inf_{\overline{\boldsymbol{\gamma}}\in\overline{\boldsymbol{\Gamma}}_{(L,p,\boldsymbol{s})}}\|\boldsymbol{\gamma}^\star-\overline{\boldsymbol{\gamma}}\|_{2,P_W}^2+\frac{\Delta(L,p,\boldsymbol{s})}{n}+\frac{V_{(L,p,\boldsymbol{s})}}{n}\log n\right]\\ &\le C_\kappa\left[\inf_{\overline{\boldsymbol{\gamma}}\in\overline{\boldsymbol{\Gamma}}_{(L,p,\boldsymbol{s})}}\|\boldsymbol{\gamma}^\star-\overline{\boldsymbol{\gamma}}\|_\infty^2+\frac{\Delta(L,p,\boldsymbol{s})}{n}+\frac{V_{(L,p,\boldsymbol{s})}}{n}\log n\right]\\ &\le C_{\kappa,k,\mathbf{d},\mathbf{t},\boldsymbol{\alpha},K}\phi_n\left[1+(\log n)^2+(\log n)^4\right]\\ &\le C_{\kappa,k,\mathbf{d},\mathbf{t},\boldsymbol{\alpha},K}\phi_n\left(\log n\right)^4.\end{aligned}$$

We complete the proof by taking the supremum over $\mathcal{F}_{[v_-,v_+]}(k,\mathbf{d},\mathbf{t},\boldsymbol{\alpha},K)$.

A.7 Proof of Corollary 6.

Proof. We note that the collection of models $\left\{\overline{\boldsymbol{\Gamma}}_m,\ m\in\mathcal{M}\right\}$ satisfies Assumption 1 with $V_m=|m|+1$. By Lemma 7, the associated weights $\Delta(m)$ satisfy inequality (7) with $\Sigma\le 1+\pi^2/6$. Moreover, for each $m\in\mathcal{M}$, the

countable subset $\mathbf{\Gamma}_m$ is dense in $\overline{\mathbf{\Gamma}}_m$ for the topology of pointwise convergence so that $h(R^\star, \mathscr{Q}_m) = h(R^\star, \overline{\mathscr{Q}}_m)$.

We apply (13) and derive that whatever the distribution of W, the resulted estimator $R_{\widehat{\boldsymbol{\gamma}}}$ satisfies

$$\mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}})\right] \leq c_2(c_3+\Sigma)(\mathcal{B}_o \wedge \mathcal{B}_c), \tag{82}$$

where

$$\mathcal{B}_o = \inf_{m\in\mathcal{M}_o}\left[h^2(R_{\boldsymbol{\gamma}^\star}, \overline{\mathscr{Q}}_m) + \frac{2\log(1+|m|)}{n} + \frac{|m|+1}{n}\left[1+\log_+\left(\frac{n}{|m|+1}\right)\right]\right],$$

$$\mathcal{B}_c = \inf_{m\in\mathcal{M}\setminus\mathcal{M}_o}\left[h^2(R_{\boldsymbol{\gamma}^\star}, \overline{\mathscr{Q}}_m) + \frac{|m|}{n}\log\left(\frac{2ep}{|m|}\right) + \frac{|m|+1}{n}\left[1+\log_+\left(\frac{n}{|m|+1}\right)\right]\right].$$

For $\mathcal{B}_o$, we observe that

$$\begin{aligned}\mathcal{B}_o &\leq \inf_{m\in\mathcal{M}_o}\left[h^2(R_{\boldsymbol{\gamma}^\star}, \overline{\mathscr{Q}}_m) + \frac{|m|+1}{n} + \frac{|m|+1}{n}\left[1+\log_+\left(\frac{n}{|m|+1}\right)\right]\right] \\ &\leq 2\inf_{m\in\mathcal{M}_o}\left[h^2(R_{\boldsymbol{\gamma}^\star}, \overline{\mathscr{Q}}_m) + \frac{|m|+1}{n}\left[1+\log_+\left(\frac{n}{|m|+1}\right)\right]\right] = 2\mathscr{B}_o. \end{aligned} \tag{83}$$

We also note that function $f(x) = x\log(2ep/x)$ is increasing on $(0, 2p]$. Therefore, for $\mathcal{B}_c$ we have

$$\begin{aligned}\mathcal{B}_c &\leq \inf_{m\in\mathcal{M}\setminus\mathcal{M}_o}\left[h^2(R_{\boldsymbol{\gamma}^\star}, \overline{\mathscr{Q}}_m) + \frac{|m|+1}{n}\left[1+\log\left(\frac{2ep}{|m|+1}\right)+\log_+\left(\frac{n}{|m|+1}\right)\right]\right] \\ &\leq 2\inf_{m\in\mathcal{M}\setminus\mathcal{M}_o}\left[h^2(R_{\boldsymbol{\gamma}^\star}, \overline{\mathscr{Q}}_m) + \frac{|m|+1}{n}\left[1+\log\left(\frac{(2p)\vee n}{|m|+1}\right)\right]\right]. \end{aligned} \tag{84}$$

Moreover, we note that for any $m \in \mathcal{M}_o$,

$$\log_+\left(\frac{n}{|m|+1}\right) \leq \log\left(\frac{(2p)\vee n}{|m|+1}\right). \tag{85}$$

Combining (82), (83), (84) and (85), we have

$$\mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}})\right] \leq 2c_2(c_3+\Sigma)(\mathscr{B}_o \wedge \mathscr{B}_c),$$

which concludes the proof. □

Appendix B. Proofs of lemmas

B.1 Proof of Lemma 1.

Proof. Let us first prove $\boldsymbol{S}^{\mathcal{B},d}_{(\boldsymbol{s},r)}$ is dense in $\overline{\boldsymbol{S}}^{\mathcal{B},d}_{(\boldsymbol{s},r)}$ with respect to the supremum norm. By the definition of dense with respect to the supremum norm, it is enough to show for any $\boldsymbol{\gamma} \in \overline{\boldsymbol{S}}^{\mathcal{B},d}_{(\boldsymbol{s},r)}$, there exists a sequence of functions $\boldsymbol{\gamma}_l \in \boldsymbol{S}^{\mathcal{B},d}_{(\boldsymbol{s},r)}$, $l \in \mathbb{N}$ such that $\lim_{l\to+\infty}\|\boldsymbol{\gamma}_l - \boldsymbol{\gamma}\|_\infty = 0$.

For any $\boldsymbol{w} \in [0,1]^d$, there is a vector $(k_1, \ldots, k_d) \in \Psi(s_1) \times \cdots \times \Psi(s_d)$ such that $\boldsymbol{w} \in \prod_{j=1}^d I_j(k_j)$. Without loss of generality, we only need to show for any function $\widetilde{\boldsymbol{\gamma}}$ on $\prod_{j=1}^d I_j(k_j)$ of the form

$$\widetilde{\boldsymbol{\gamma}}(\boldsymbol{w}) = \sum_{(r_1,\ldots,r_d)\in\{0,\ldots,r\}^d} \widetilde{\gamma}_{(r_1,\ldots,r_d)} \prod_{j=1}^d {w_j}^{r_j}, \tag{86}$$

where $\widetilde{\gamma}_{(r_1,\ldots,r_d)} \in \mathbb{R}$, for all $0 \le r_j \le r$, $1 \le j \le d$, there is a sequence of functions $\{\widetilde{\boldsymbol{\gamma}}_l\}_{l\in\mathbb{N}}$ on $\prod_{j=1}^d I_j(k_j)$ of the form

$$\widetilde{\boldsymbol{\gamma}}_l(\boldsymbol{w}) = \sum_{(r_1,\ldots,r_d)\in\{0,\ldots,r\}^d} \widetilde{\gamma}^l_{(r_1,\ldots,r_d)} \prod_{j=1}^d {w_j}^{r_j}, \tag{87}$$

with $\widetilde{\gamma}^l_{(r_1,\ldots,r_d)} \in \mathbb{Q}$, for all $0 \le r_j \le r$, $1 \le j \le d$ and $l \in \mathbb{N}$ such that $\lim_{l\to+\infty} \sup_{\boldsymbol{w}\in\prod_{j=1}^d I_j(k_j)} |\widetilde{\boldsymbol{\gamma}}_l(\boldsymbol{w}) - \widetilde{\boldsymbol{\gamma}}(\boldsymbol{w})| = 0$.

In fact, since $\mathbb{Q}$ is dense in $\mathbb{R}$, for all $\widetilde{\gamma}_{(r_1,\ldots,r_d)} \in \mathbb{R}$ with $(r_1, \ldots, r_d) \in \{0, \ldots, r\}^d$ and all $\epsilon > 0$, there is a sequence of rational numbers $\widetilde{\gamma}^l_{(r_1,\ldots,r_d)} \in \mathbb{Q}$ and a $N_\epsilon > 0$ such that for all $l \ge N_\epsilon$,

$$\left|\widetilde{\gamma}_{(r_1,\ldots,r_d)} - \widetilde{\gamma}^l_{(r_1,\ldots,r_d)}\right| < \frac{\epsilon}{(r+1)^d}.$$

Hence, for any $\widetilde{\boldsymbol{\gamma}}$ defined by (86) and all $\epsilon > 0$, there exists a $N_\epsilon > 0$ and a sequence of functions $\{\widetilde{\boldsymbol{\gamma}}_l\}_{l\in\mathbb{N}}$ defined by (87) such that for all $l \ge N_\epsilon$,

$$\begin{aligned}
\sup_{\boldsymbol{w}\in\prod_{j=1}^d I_j(k_j)} |\widetilde{\boldsymbol{\gamma}}_l(\boldsymbol{w}) - \widetilde{\boldsymbol{\gamma}}(\boldsymbol{w})| &\le \left| \sum_{(r_1,\ldots,r_d)\in\{0,\ldots,r\}^d} \left(\widetilde{\gamma}_{(r_1,\ldots,r_d)} - \widetilde{\gamma}^l_{(r_1,\ldots,r_d)}\right) \right| \\
&\le \sum_{(r_1,\ldots,r_d)\in\{0,\ldots,r\}^d} \left|\widetilde{\gamma}_{(r_1,\ldots,r_d)} - \widetilde{\gamma}^l_{(r_1,\ldots,r_d)}\right| \\
&< \sum_{(r_1,\ldots,r_d)\in\{0,\ldots,r\}^d} \frac{\epsilon}{(r+1)^d} \le \epsilon.
\end{aligned}$$

The conclusion then follows by the definition of limit.

To prove that $\boldsymbol{\Gamma}^{\mathcal{B},d}_{(\boldsymbol{s},r)}$ is dense in $\overline{\boldsymbol{\Gamma}}^{\mathcal{B},d}_{(\boldsymbol{s},r)}$ with respect to the supremum norm, it is enough to note that for any $f \in \boldsymbol{S}^{\mathcal{B},d}_{(\boldsymbol{s},r)}$ and $g \in \overline{\boldsymbol{S}}^{\mathcal{B},d}_{(\boldsymbol{s},r)}$

$$\|(f \vee v_-) \wedge v_+ - (g \vee v_-) \wedge v_+\|_\infty \le \|f - g\|_\infty.$$

□

B.2 Proof of Lemma 2.

Proof. For any $D \in \mathbb{N}^*$, let M_D^d stand for the set of partitions which divide $[0,1]^d$ into D hyperrectangles. Since $\cup_{\boldsymbol{s}\in\mathbb{N}^d} M_{\boldsymbol{s}}^{\mathcal{B},d} \subset \cup_{D\in\mathbb{N}^*} M_D^d$, we have

$$\sum_{(\boldsymbol{s},r)\in\mathcal{M}} \exp\left[-\log(8d)\prod_{j=1}^{d} 2^{s_j} - r\right] \leq \sum_{r\in\mathbb{N}} \sum_{D\in\mathbb{N}^*} \sum_{\pi\in M_D^d} e^{-\log(8d)|\pi|-r}, \tag{88}$$

where $|\pi|$ denotes the cardinality of hyperrectangles given by the partition π of $[0,1]^d$.

By the proof of Proposition 5 in Akakpo (2012), a partition over $[0,1]^d$ into D hyperrectangles addresses to choosing a vector $(l_1,\ldots,l_{D-1}) \in \{1,\ldots,d\}^{D-1}$ for the partition directions and growing a binary tree with root $[0,1]^d$ and D leaves. The number of partitions belonging to M_D^d satisfies $|M_D^d| \leq (4d)^D$. Therefore, we derive from (88) that

$$\begin{aligned}\sum_{(\boldsymbol{s},r)\in\mathcal{M}} \exp\left[-\log(8d)\prod_{j=1}^{d} 2^{s_j} - r\right] &\leq \sum_{r\in\mathbb{N}} e^{-r}\left(\sum_{D\in\mathbb{N}^*} (4d)^D (8d)^{-D}\right)\\ &\leq \sum_{r\in\mathbb{N}} e^{-r} = \frac{e}{e-1}.\end{aligned}$$

□

B.3 Proof of Lemma 4.

Proof. It is equivalent to prove

$$\sum_{(\boldsymbol{s},t,r)\in\mathbb{N}^d\times\mathbb{N}^*\times\mathbb{N}} e^{-\Delta(\boldsymbol{s},t,r)} \leq \frac{e}{e-1},$$

where

$$\Delta(\boldsymbol{s},t,r) = 3\log 2\left(\sum_{j=1}^{d} 2^{s_j}\right) + r + t.$$

For $(D_1,\ldots,D_d) \in (\mathbb{N}^*)^d$, let $M_{D_j}^1$ represent the set of partitions which divide $[0,1]$ into D_j subintervals. Recall that $M_s^{\mathcal{B},1}$ denotes the dyadic partition of $[0,1]$ into 2^s subintervals, hence we have for any $j \in \{1,\ldots,d\}$,

$\cup_{s_j \in \mathbb{N}} M_{s_j}^{\mathcal{B},1} \subset \cup_{D_j \in \mathbb{N}^*} M_{D_j}^1$. As an immediate consequence,

$$\begin{aligned}
&\sum_{(\boldsymbol{s},t,r)\in\mathbb{N}^d\times\mathbb{N}^*\times\mathbb{N}} e^{-\Delta(\boldsymbol{s},t,r)} \\
&= \sum_{(\boldsymbol{s},t,r)\in\mathbb{N}^d\times\mathbb{N}^*\times\mathbb{N}} \exp\left[-t - \sum_{j=1}^{d} 2^{s_j}\log 8 - r\right] \\
&\leq \sum_{r\in\mathbb{N}} e^{-r}\left[\prod_{j=1}^{d}\left(\sum_{D_j\in\mathbb{N}^*}\sum_{\pi_j\in M_{D_j}^1} e^{-|\pi_j|\log 8}\right)\right]\left(\sum_{t\in\mathbb{N}^*} e^{-t}\right), \qquad (89)
\end{aligned}$$

where $|\pi_j|$ denotes the cardinality of segments given by the partition π_j. Moreover, as we have mentioned in the proof of Lemma 2, it follows from Proposition 5 in Akakpo (2012) that $|M_{D_j}^1| \leq 4^{D_j}$ for $j \in \{1, \ldots, d\}$. Therefore, we derive from (89) that

$$\begin{aligned}
\sum_{(\boldsymbol{s},t,r)\in\mathbb{N}^d\times\mathbb{N}^*\times\mathbb{N}} e^{-\Delta(\boldsymbol{s},t,r)} &\leq \sum_{r\in\mathbb{N}} e^{-r}\left(\sum_{D\in\mathbb{N}^*} 4^D 8^{-D}\right)^d\left(\sum_{t\in\mathbb{N}^*} e^{-t}\right) \\
&\leq \sum_{r\in\mathbb{N}} e^{-r} \leq \frac{e}{e-1}.
\end{aligned}$$

□

B.4 Proof of Lemma 5.

Proof.

$$\begin{aligned}
\sum_{(\boldsymbol{t},r)\in(\mathbb{N}^*)^l\times\mathbb{N}} e^{-\Delta(\boldsymbol{t},r)} &= \sum_{(\boldsymbol{t},r)\in(\mathbb{N}^*)^l\times\mathbb{N}} \exp\left[-\sum_{j=1}^{l} t_j - r\right] \\
&\leq \sum_{r\in\mathbb{N}} e^{-r}\left(\sum_{\boldsymbol{t}\in(\mathbb{N}^*)^l} e^{-\sum_{j=1}^{l} t_j}\right) \\
&\leq \sum_{r\in\mathbb{N}} e^{-r}\left(\sum_{t\in\mathbb{N}^*} e^{-t}\right)^l \\
&\leq \frac{e}{e-1}.
\end{aligned}$$

□

B.5 Proof of Lemma 6. We hereby introduce a combinatorial result given by Proposition 2.5 of Massart (2007): for all integers $|m|$ and p with $1 \leq$

$|m| \leq p$,

$$\sum_{k=0}^{|m|} \binom{p}{k} \leq \left(\frac{ep}{|m|}\right)^{|m|}. \tag{90}$$

Proof. First, we note that

$$\begin{aligned}
&\sum_{(L,p,\boldsymbol{s})\in(\mathbb{N}^*)^2\times\{0,1\}^{\overline{p}}} e^{-\Delta(L,p,\boldsymbol{s})} \\
&= \sum_{L\in\mathbb{N}^*} e^{-L}\left[\sum_{p\in\mathbb{N}^*} e^{-p}\left(1+\sum_{\boldsymbol{s}\in\{0,1\}^{\overline{p}}\setminus\{0\}^{\overline{p}}} \exp\left[-\|\boldsymbol{s}\|_0 \log\left(\frac{2e\overline{p}}{\|\boldsymbol{s}\|_0}\right)\right]\right)\right] \\
&\leq \sum_{L\in\mathbb{N}^*} e^{-L}\left\{\sum_{p\in\mathbb{N}^*} e^{-p}\left[1+\sum_{s=1}^{\overline{p}}\binom{\overline{p}}{s}\exp\left(-s\log\left(\frac{2e\overline{p}}{s}\right)\right)\right]\right\}. \qquad (91)
\end{aligned}$$

By (90), we know for any $1 \leq s \leq \overline{p}$,

$$\binom{\overline{p}}{s} \leq \sum_{h=0}^{s}\binom{\overline{p}}{h} \leq \left(\frac{e\overline{p}}{s}\right)^{s}. \tag{92}$$

Plugging (92) into (91), we obtain

$$\begin{aligned}
&\sum_{(L,p,\boldsymbol{s})\in(\mathbb{N}^*)^2\times\{0,1\}^{\overline{p}}} e^{-\Delta(L,p,\boldsymbol{s})} \\
&\leq \sum_{L\in\mathbb{N}^*} e^{-L}\left\{\sum_{p\in\mathbb{N}^*} e^{-p}\left[1+\sum_{s=1}^{\overline{p}}\left(\frac{e\overline{p}}{s}\right)^{s}\left(\frac{2e\overline{p}}{s}\right)^{-s}\right]\right\} \\
&\leq \sum_{L\in\mathbb{N}^*} e^{-L}\left[\sum_{p\in\mathbb{N}^*} e^{-p}\left(\sum_{s=0}^{\overline{p}} 2^{-s}\right)\right] \\
&\leq \sum_{L\in\mathbb{N}^*} e^{-L}\left[\sum_{p\in\mathbb{N}^*} e^{-p}\left(\sum_{s=0}^{+\infty} 2^{-s}\right)\right] \leq 2.
\end{aligned}$$

□

B.6 Proof of Lemma 7.

Proof. By (90), we derive that

$$\begin{aligned}
\Sigma = \sum_{m\in\mathcal{M}} e^{-\Delta(m)} &= \sum_{m\in\mathcal{M}_o} e^{-\Delta(m)} + \sum_{m\in\mathcal{M}\setminus\mathcal{M}_o} e^{-\Delta(m)} \\
&\le \sum_{d=0}^{p} \frac{1}{(1+d)^2} + \sum_{|m|=1}^{p} \binom{p}{|m|} \exp\left[-|m|\log\left(\frac{2ep}{|m|}\right)\right] \\
&\le \sum_{k=1}^{+\infty} \frac{1}{k^2} + \sum_{|m|=1}^{p} \left(\frac{ep}{|m|}\right)^{|m|} \exp\left[-|m|\log\left(\frac{2ep}{|m|}\right)\right] \\
&\le \frac{\pi^2}{6} + \sum_{|m|=1}^{+\infty} 2^{-|m|} \\
&\le \frac{\pi^2}{6} + 1.
\end{aligned}$$

□

Appendix C. Proofs of VC dimensions

The proofs in this section are inspired by the proof of Theorem 7 in Barlett et al. (2019). We first introduce three results which we shall use later for deriving the VC dimension bounds. The first one is the result of Lemma 1 in Barlett et al. (1998).

Lemma 12. *Suppose $f_1(\cdot), f_2(\cdot), \ldots, f_T(\cdot)$ are fixed polynomials of degree at most d in $s \le T$ variables. Define*

$$N := \left|\left\{(\operatorname{sgn}(f_1(a)), \ldots, \operatorname{sgn}(f_T(a)))\,,\ a \in \mathbb{R}^s\right\}\right|,$$

i.e., N is the number of distinct sign vectors generated by varying $a \in \mathbb{R}^s$. Then we have $N \le 2(2edT/s)^s$.

The second Lemma is the weighted AM-GM Inequality.

Lemma 13 (Weighted AM-GM Inequality)**.** *If $0 \le c_i \in \mathbb{R}$ and $0 \le \lambda_i \in \mathbb{R}$ for all $i = 1, \ldots, K$ such that $\sum_{i=1}^{K} \lambda_i = 1$, then*

$$\prod_{i=1}^{K} c_i^{\lambda_i} \le \sum_{i=1}^{K} \lambda_i c_i.$$

The third result comes from the Lemma 18 of Barlett et al. (2019).

Lemma 14. *Suppose that $2^m \le 2^t (mr/w)^w$ for some $r \ge 16$ and $m \ge w \ge t \ge 0$. Then, $m \le t + w\log_2(2r\log_2 r)$.*

C.1 Proof of Proposition 1.

Proof. For a given $r \in \mathbb{N}$, $t \in \mathbb{N}^*$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_d) \in (M^{\mathcal{B},1})^d$, we define $\widetilde{\boldsymbol{\Gamma}}^A_{(\boldsymbol{\pi},t,r)}$ the collection of all the functions on $\mathscr{W} = [0,1]^d$ of the form

$$\boldsymbol{\gamma}(\boldsymbol{w}) = f\left[(g(\boldsymbol{w}) \vee 0) \wedge 1\right], \quad \text{for all } \boldsymbol{w} = (w_1, \ldots, w_d) \in [0,1]^d,$$

where $g(\boldsymbol{w}) = \sum_{j=1}^d g_j(w_j)$ with $g_j \in \overline{\boldsymbol{S}}^{\mathcal{B},1}_{(\pi_j,r)}$, for all $j \in \{1, \ldots, d\}$ and $f \in \overline{\boldsymbol{S}}^{\mathcal{H},1}_{(t,r)}$. The class of functions $\overline{\boldsymbol{S}}^{\mathcal{B},1}_{(\pi_j,r)}$ has been defined in Section 4 and $\overline{\boldsymbol{S}}^{\mathcal{H},1}_{(t,r)}$ in Section 5. Let $V^{\widetilde{A}}_{(\boldsymbol{\pi},t,r)}$ denote the VC dimension of $\widetilde{\boldsymbol{\Gamma}}^A_{(\boldsymbol{\pi},t,r)}$. We first prove the conclusion holds for $\widetilde{\boldsymbol{\Gamma}}^A_{(\boldsymbol{\pi},t,r)}$, i.e.

$$V^{\widetilde{A}}_{(\boldsymbol{\pi},t,r)} \leq 2 + \left[t(r+1) + 2\sum_{j=1}^d |\pi_j|(r+1)\right] \log_2\left[4eU \log_2(2eU)\right],$$

where $U = t+r+2$. Then, by rewriting $\overline{\boldsymbol{\Gamma}}^A_{(\boldsymbol{\pi},t,r)} = \left\{(\boldsymbol{\gamma} \vee v_-) \wedge v_+,\ \boldsymbol{\gamma} \in \widetilde{\boldsymbol{\Gamma}}^A_{(\boldsymbol{\pi},t,r)}\right\}$, the conclusion also holds for $\overline{\boldsymbol{\Gamma}}^A_{(\boldsymbol{\pi},t,r)}$ according to the properties of VC-subgraph we introduced in Section 3.

Recall that $\overline{\boldsymbol{S}}^{\mathcal{B},1}_{(\pi_j,r)}$ is a $|\pi_j|(r+1)$ dimensional vector space for any $j \in \{1, \ldots, d\}$ and $\overline{\boldsymbol{S}}^{\mathcal{H},1}_{(t,r)}$ is a $t(r+1)$ dimensional vector space. Therefore, any element belonging to $\widetilde{\boldsymbol{\Gamma}}^A_{(\boldsymbol{\pi},t,r)}$ is determined by a vector of real numbers $a \in \mathbb{R}^s$ with $s = \left(t + \sum_{j=1}^d |\pi_j|\right)(r+1)$ which we call parameters in the sequel. We denote g_a the function $g(\boldsymbol{w}) = \sum_{j=1}^d g_j(w_j)$ and f_a the function in $\widetilde{\boldsymbol{\Gamma}}^A_{(\boldsymbol{\pi},t,r)}$ induced by the parameters vector $a \in \mathbb{R}^s$ hence we have $\widetilde{\boldsymbol{\Gamma}}^A_{(\boldsymbol{\pi},t,r)} = \{f_a,\ a \in \mathbb{R}^s\}$. Given a fixed point $\boldsymbol{w}$ on $\mathscr{W}$, for any $a \in \mathbb{R}^s$, we denote $h_{\boldsymbol{w}}(a) = f_a(\boldsymbol{w})$ and $h'_{\boldsymbol{w}}(a) = g_a(\boldsymbol{w})$.

We take m fixed points $(\boldsymbol{w}_1, v_1), \ldots, (\boldsymbol{w}_m, v_m) \in \mathscr{W} \times \mathbb{R}$, where for each $i \in \{1, \ldots, m\}$, $\boldsymbol{w}_i = (w_i^1, \ldots, w_i^d) \in [0,1]^d$. We first derive a bound for the total number of signs patterns given fixed $(\boldsymbol{w}_1, v_1), \ldots, (\boldsymbol{w}_m, v_m) \in \mathscr{W} \times \mathbb{R}$, i.e.

$$N(m) = \Big|\left\{(\operatorname{sgn}(h_{\boldsymbol{w}_1}(a) - v_1), \ldots, \operatorname{sgn}(h_{\boldsymbol{w}_m}(a) - v_m)),\ a \in \mathbb{R}^s\right\}\Big|.$$

The idea is to construct a special partition $\mathcal{S}$ of $\mathbb{R}^s$ where within each region $S \in \mathcal{S}$ the functions $h_{\boldsymbol{w}_i}(a) - v_i$, $i \in \{1, \ldots, m\}$ are all fixed polynomials of a with a bounded degree.

We start with $\mathcal{S}_0 = \{\mathbb{R}^s\}$. For any $i \in \{1, \ldots, m\}$, we note that $h'_{\boldsymbol{w}_i}(a)$ is a fixed polynomial depending on at most $\sum_{j=1}^d |\pi_j|(r+1)$ variables with the total degree no more than 1. We recall that for a given $t \in \mathbb{N}^*$, $M_t^{\mathcal{H},1}$ defined in Section 5 is the regular partition of $[0,1]$ into t subintervals. Let

$\{b_1, \dots, b_{t-1}\}$ be the breakpoints on the interval $(0,1)$ given by $M_t^{\mathcal{H},1}$ and denote $b_0 = 0$, $b_t = 1$. Applying Lemma 12 to the collection of polynomials

$$\mathcal{C} = \left\{ h'_{\boldsymbol{w}_i}(a) - b_l, i \in \{1, \dots, m\}, l \in \{0, \dots, t\} \right\},$$

we know that when a varies in $\mathbb{R}^s$, it attains at most

$$N_1 := 2 \left(\frac{2em(t+1)}{\sum_{j=1}^d |\pi_j|(r+1)} \right)^{\sum_{j=1}^d |\pi_j|(r+1)}$$

distinct signs patterns. Therefore, one can partition $\mathbb{R}^s$ into N_1 pieces with the refined partition $\mathcal{S}_1 = \{S_1, \dots, S_{N_1}\}$ such that all the polynomials in $\mathcal{C}$ have fixed signs within each region $S \in \mathcal{S}_1$. For any $S \in \mathcal{S}_1$ and any $i \in \{1, \dots, m\}$, when a varies in S, $h_{\boldsymbol{w}_i}(a)$ is a fixed polynomial of at most $(t + \sum_{j=1}^d |\pi_j|)(r+1)$ variables with the total degree no more than $r+1$. Hence by Lemma 12 again, on each $S \in \mathcal{S}_1$,

$$\{(\operatorname{sgn}(h_{\boldsymbol{w}_1}(a) - v_1), \dots, \operatorname{sgn}(h_{\boldsymbol{w}_m}(a) - v_m)),\ a \in S\}$$

has at most

$$N_2 := 2 \left(\frac{2em(r+1)}{(t + \sum_{j=1}^d |\pi_j|)(r+1)} \right)^{(t+\sum_{j=1}^d |\pi_j|)(r+1)}$$

distinct signs patterns. We intersect all these regions with $S \in \mathcal{S}_1$ which yields a refined partition $\mathcal{S}_2 = \{S_1, \dots, S_{N_1 N_2}\}$ over $\mathbb{R}^s$ with at most $N_1 N_2$ pieces such that within each region $S \in \mathcal{S}_2$,

$$(\operatorname{sgn}(h_{\boldsymbol{w}_1}(a) - v_1), \dots, \operatorname{sgn}(h_{\boldsymbol{w}_m}(a) - v_m))$$

have unchanged signs patterns when a varies in S. We denote

$$\lambda_1 = \frac{\sum_{j=1}^d |\pi_j|(r+1)}{t(r+1) + 2\sum_{j=1}^d |\pi_j|(r+1)}, \quad \lambda_2 = \frac{(t + \sum_{j=1}^d |\pi_j|)(r+1)}{t(r+1) + 2\sum_{j=1}^d |\pi_j|(r+1)},$$

$$c_1 = \frac{2em(t+1)}{\sum_{j=1}^d |\pi_j|(r+1)}, \quad c_2 = \frac{2em(r+1)}{(t + \sum_{j=1}^d |\pi_j|)(r+1)}.$$

For any arbitrarily chosen m points $(\boldsymbol{w}_1, v_1), \dots, (\boldsymbol{w}_m, v_m) \in \mathscr{W} \times \mathbb{R}$, we have

$$\begin{aligned} N(m) &\le \sum_{k=1}^{N_1 N_2} |\{(\operatorname{sgn}(h_{\boldsymbol{w}_1}(a) - v_1), \dots, \operatorname{sgn}(h_{\boldsymbol{w}_m}(a) - v_m)),\ a \in S_k\}| \\ &\le N_1 N_2 \le 4 \left(c_1^{\lambda_1} c_2^{\lambda_2} \right)^{t(r+1) + 2\sum_{j=1}^d |\pi_j|(r+1)}. \end{aligned} \tag{93}$$

Applying Lemma 13 to (93), we derive that

$$\begin{aligned} N(m) &\le N_1 N_2 \le 4 \left(\lambda_1 c_1 + \lambda_2 c_2 \right)^{t(r+1) + 2\sum_{j=1}^d |\pi_j|(r+1)} \\ &\le 4 \left(\frac{2em(t+r+2)}{t(r+1) + 2\sum_{j=1}^d |\pi_j|(r+1)} \right)^{t(r+1) + 2\sum_{j=1}^d |\pi_j|(r+1)}. \end{aligned} \tag{94}$$

From the definition of VC-dimension together with (94),

$$2^{V^{\widetilde{A}}_{(\boldsymbol{\pi},t,r)}} = N\left[V^{\widetilde{A}}_{(\boldsymbol{\pi},t,r)}\right] \leq 4\left(\frac{2e(t+r+2)V^{\widetilde{A}}_{(\boldsymbol{\pi},t,r)}}{t(r+1)+2\sum_{j=1}^{d}|\pi_j|(r+1)}\right)^{t(r+1)+2\sum_{j=1}^{d}|\pi_j|(r+1)}.$$

We denote $U = t + r + 2$. Since $r \in \mathbb{N}$ and $t \in \mathbb{N}^*$, we have $U \geq 3$ and $2eU \geq 16$. We then can apply Lemma 14 and obtain

$$V^{\widetilde{A}}_{(\boldsymbol{\pi},t,r)} \leq 2 + \left[t(r+1) + 2\sum_{j=1}^{d}|\pi_j|(r+1)\right]\log_2\left[4eU\log_2(2eU)\right].$$

The conclusion finally follows by $V^{A}_{(\boldsymbol{\pi},t,r)} \leq V^{\widetilde{A}}_{(\boldsymbol{\pi},t,r)}$. □

C.2 Proof of Proposition 2.

Proof. For a given $r \in \mathbb{N}$ and $\boldsymbol{t} = (t_1, \ldots, t_l) \in (\mathbb{N}^*)^l$, we define $\widetilde{\boldsymbol{\Gamma}}^{M}_{(\boldsymbol{t},r)}$ the collection of all the functions $\boldsymbol{\gamma}$ on $[0,1]^d$ of the form

$$\boldsymbol{\gamma}(\boldsymbol{w}) = f\left(g_1(\boldsymbol{w}), \ldots, g_l(\boldsymbol{w})\right), \text{ for all } \boldsymbol{w} \in [0,1]^d \tag{95}$$

where $f \in \overline{\boldsymbol{S}}^{\mathcal{H},l}_{(\boldsymbol{t},r)}$, $g_j(\boldsymbol{w}) = \left[\left(\left(\langle a_j, \boldsymbol{w}\rangle + 1\right)/2\right) \vee 0\right] \wedge 1$ with $a_j \in \mathbb{R}^d$ for all $j \in \{1, \ldots, l\}$. We denote $V^{\widetilde{M}}_{(\boldsymbol{t},r)}$ the VC dimension of the class of functions $\widetilde{\boldsymbol{\Gamma}}^{M}_{(\boldsymbol{t},r)}$. We first prove

$$V^{\widetilde{M}}_{(\boldsymbol{t},r)} \leq 2 + \left[2ld + \left(\prod_{j=1}^{l} t_j\right)(r+1)^l\right]\log_2\left[4eU\log_2(2eU)\right],$$

where $U = \sum_{j=1}^{l} t_j + lr + l + 1$.

Let us recall that by the definition of $\overline{\boldsymbol{S}}^{\mathcal{H},l}_{(\boldsymbol{t},r)}$ in Section 5 and (95), any function belonging to $\widetilde{\boldsymbol{\Gamma}}^{M}_{(\boldsymbol{t},r)}$ is determined by a vector of real numbers $a \in \mathbb{R}^s$ with $s = ld + \left(\prod_{j=1}^{l} t_j\right)(r+1)^l$ which we call parameters in the sequel. We denote f_a the function in $\widetilde{\boldsymbol{\Gamma}}^{M}_{(\boldsymbol{t},r)}$ induced by the parameters vector $a \in \mathbb{R}^s$ hence we can rewrite $\widetilde{\boldsymbol{\Gamma}}^{M}_{(\boldsymbol{t},r)} = \{f_a,\ a \in \mathbb{R}^s\}$. Given a fixed point $\boldsymbol{w}$ on $\mathscr{W}$, for any $a \in \mathbb{R}^s$, we denote $h_{\boldsymbol{w}}(a) = f_a(\boldsymbol{w})$.

We start with fixing m points $(\boldsymbol{w}_1, v_1), \ldots, (\boldsymbol{w}_m, v_m) \in \mathscr{W} \times \mathbb{R}$. Provided m fixed points $(\boldsymbol{w}_1, v_1), \ldots, (\boldsymbol{w}_m, v_m) \in \mathscr{W} \times \mathbb{R}$, we first bound the total number of signs patterns, i.e.

$$N(m) = \Big|\left\{\left(\operatorname{sgn}(h_{\boldsymbol{w}_1}(a) - v_1), \ldots, \operatorname{sgn}(h_{\boldsymbol{w}_m}(a) - v_m)\right),\ a \in \mathbb{R}^s\right\}\Big|.$$

The idea is similar to the proof of Proposition 1 which is to construct a special partition $\mathcal{S}$ of $\mathbb{R}^s$ such that within each region $S \in \mathcal{S}$, the functions $h_{\boldsymbol{w}_i}(a) - v_i$, for $i \in \{1, \ldots, m\}$ are all fixed polynomials of a with a bounded degree. Therefore, we can conclude by applying Lemma 12.

We initialise our partition of $\mathbb{R}^s$ with $\mathcal{S}_0 = \{\mathbb{R}^s\}$. We recall that for a given $\boldsymbol{t} = (t_1, \dots, t_l) \in (\mathbb{N}^*)^l$, $M_{\boldsymbol{t}}^{\mathcal{H},l}$ defined in Section 5 is the partition of $[0,1]^l$, where in all the directions $j \in \{1, \dots, l\}$, the interval $[0,1]$ is divided into t_j regular subintervals. Let $\{b_1^j, \dots, b_{t_j-1}^j\}$ be the breakpoints on the interval $(0,1)$ in the j-th direction given by the partition $M_{\boldsymbol{t}}^{\mathcal{H},l}$ and denote $b_0^j = 0$, $b_{t_j}^j = 1$ for all $j \in \{1, \dots, l\}$. We consider the collection of polynomials

$$\mathcal{C} = \left\{ \frac{1}{2} \left(\langle a_j, \boldsymbol{w}_i \rangle + 1 \right) - b_k^j,\ i \in \{1, \dots, m\},\ j \in \{1, \dots, l\},\ k \in \{0, \dots, t_j\} \right\}.$$

Since all the functions in $\mathcal{C}$ can be written as a fixed polynomial of degree no more than 1 in ld variables of a, $\mathcal{C}$ attains at most

$$N_1 := 2 \left(\frac{2em \sum_{j=1}^l (t_j + 1)}{ld} \right)^{ld}$$

distinct signs patterns when a varies in $\mathbb{R}^s$ according to Lemma 12. Therefore, we partition $\mathbb{R}^s$ into N_1 pieces with the refined partition $\mathcal{S}_1 = \{S_1, \dots, S_{N_1}\}$ such that within each region $S \in \mathcal{S}_1$, all the polynomials in $\mathcal{C}$ have fixed signs when a varies in S. Now we consider on each $S \in \mathcal{S}_1$, for any $i \in \{1, \dots, m\}$, $h_{\boldsymbol{w}_i}(a)$ with $a \in S$ is a fixed polynomial of at most $ld + \left(\prod_{j=1}^l t_j \right) (r+1)^l$ variables with the total degree no more than $lr + 1$. Hence by Lemma 12 again, on each $S \in \mathcal{S}_1$,

$$\{ (\operatorname{sgn}(h_{\boldsymbol{w}_1}(a) - v_1), \dots, \operatorname{sgn}(h_{\boldsymbol{w}_m}(a) - v_m)),\ a \in S \}$$

attains at most

$$N_2 := 2 \left(\frac{2em(lr+1)}{ld + \left(\prod_{j=1}^l t_j \right) (r+1)^l} \right)^{ld + \left(\prod_{j=1}^l t_j \right)(r+1)^l}$$

distinct signs patterns when a varies in S. We intersect all these regions with $S \in \mathcal{S}_1$ which yields a refined partition $\mathcal{S}_2 = \{S_1, \dots, S_{N_1 N_2}\}$ of $\mathbb{R}^s$ with at most $N_1 N_2$ pieces such that within each region, $(\operatorname{sgn}(h_{\boldsymbol{w}_1}(a) - v_1), \dots, \operatorname{sgn}(h_{\boldsymbol{w}_m}(a) - v_m))$ have unchanged signs patterns when a varies. We denote

$$\lambda_1 = \frac{ld}{2ld + \left(\prod_{j=1}^l t_j \right) (r+1)^l}, \quad \lambda_2 = \frac{ld + \left(\prod_{j=1}^l t_j \right) (r+1)^l}{2ld + \left(\prod_{j=1}^l t_j \right) (r+1)^l},$$

$$c_1 = \frac{2em \sum_{j=1}^l (t_j + 1)}{ld}, \quad c_2 = \frac{2em(lr+1)}{ld + \left(\prod_{j=1}^l t_j \right) (r+1)^l}.$$

For any arbitrarily chosen m points $(\boldsymbol{w}_1, v_1), \ldots, (\boldsymbol{w}_m, v_m) \in \mathscr{W} \times \mathbb{R}$, we have

$$N(m) \leq \sum_{k=1}^{N_1 N_2} |\left\{(\operatorname{sgn}(h_{\boldsymbol{w}_1}(a) - v_1), \ldots, \operatorname{sgn}(h_{\boldsymbol{w}_m}(a) - v_m)),\ a \in S_k\right\}|$$

$$\leq N_1 N_2 \leq 4 \left(c_1^{\lambda_1} c_2^{\lambda_2}\right)^{2ld + \left(\prod_{j=1}^{l} t_j\right)(r+1)^l}. \tag{96}$$

Applying Lemma 13 to (96), we derive that

$$\begin{aligned} N(m) &\leq N_1 N_2 \leq 4 \left(\lambda_1 c_1 + \lambda_2 c_2\right)^{2ld + \left(\prod_{j=1}^{l} t_j\right)(r+1)^l} \\ &\leq 4 \left[\frac{2em\left(\sum_{j=1}^{l} t_j + lr + l + 1\right)}{2ld + \left(\prod_{j=1}^{l} t_j\right)(r+1)^l}\right]^{2ld + \left(\prod_{j=1}^{l} t_j\right)(r+1)^l}. \end{aligned} \tag{97}$$

From the definition of VC-dimension together with (97),

$$2^{V^{\widetilde{M}}_{(\boldsymbol{t},r)}} = N\left[V^{\widetilde{M}}_{(\boldsymbol{t},r)}\right] \leq 4 \left[\frac{2e(\sum_{j=1}^{l} t_j + lr + l + 1)V^{\widetilde{M}}_{(\boldsymbol{t},r)}}{2ld + \left(\prod_{j=1}^{l} t_j\right)(r+1)^l}\right]^{2ld + \left(\prod_{j=1}^{l} t_j\right)(r+1)^l}.$$

We denote $U = \sum_{j=1}^{l} t_j + lr + l + 1$. Since $r \in \mathbb{N}$ and $\boldsymbol{t} \in (\mathbb{N}^*)^l$ with $l \in \mathbb{N}^*$, we have $U \geq 3$ and $2eU \geq 16$. We then can apply Lemma 14 and obtain

$$V^{\widetilde{M}}_{(\boldsymbol{t},r)} \leq 2 + \left[2ld + \left(\prod_{j=1}^{l} t_j\right)(r+1)^l\right] \log_2\left[4eU \log_2(2eU)\right].$$

For a given $r \in \mathbb{N}$ and $\boldsymbol{t} = (t_1, \ldots, t_l) \in (\mathbb{N}^*)^l$, we define the class of functions $\widetilde{\overline{\boldsymbol{\Gamma}}}^M_{(\boldsymbol{t},r)}$ on $\mathscr{W} = [0,1]^d$ as

$$\widetilde{\overline{\boldsymbol{\Gamma}}}^M_{(\boldsymbol{t},r)} = \left\{f(g_1, \ldots, g_l),\ f \in \overline{\boldsymbol{S}}^{\mathcal{H},l}_{(\boldsymbol{t},r)},\ g_j(\boldsymbol{w}) = \frac{\langle a_j, \boldsymbol{w}\rangle + 1}{2} \text{ with } a_j \in \mathcal{C}_d,\ j \in [l]\right\}$$

and denote $V^{\widetilde{\overline{M}}}_{(\boldsymbol{t},r)}$ the VC dimension of it. We observe that $\widetilde{\overline{\boldsymbol{\Gamma}}}^M_{(\boldsymbol{t},r)}$ is a subset of $\widetilde{\boldsymbol{\Gamma}}^M_{(\boldsymbol{t},r)}$, therefore we have $V^{\widetilde{\overline{M}}}_{(\boldsymbol{t},r)} \leq V^{\widetilde{M}}_{(\boldsymbol{t},r)}$. The conclusion finally follows by $\overline{\boldsymbol{\Gamma}}^M_{(\boldsymbol{t},r)} = \left\{(\boldsymbol{\gamma} \vee v_-) \wedge v_+,\ \boldsymbol{\gamma} \in \widetilde{\overline{\boldsymbol{\Gamma}}}^M_{(\boldsymbol{t},r)}\right\}$ so that $V^M_{(\boldsymbol{t},r)} \leq V^{\widetilde{\overline{M}}}_{(\boldsymbol{t},r)}$. □

C.3 Proof of Proposition 3.

Proof. We note that for any function $f \in \overline{\boldsymbol{S}}_{(L,p,\boldsymbol{s})}$, it is determined by the values of non-zero parameters in the weight matrices A_l and shift vectors b_l, $l \in \{0, \ldots, L\}$. For each $l \in \{0, \ldots, L\}$, we denote s_l the number of non-zero parameters in A_l and b_l and $s = \sum_{l=0}^{L} s_l$ which is exact the value of $\|\boldsymbol{s}\|_0$. Given $L \in \mathbb{N}^*$, $p \in \mathbb{N}^*$ and $\boldsymbol{s} \in \{0,1\}^{\overline{p}}$, the total number of parameters determining $f \in \overline{\boldsymbol{S}}_{(L,p,\boldsymbol{s})}$ is s. We denote f_a the function in $\overline{\boldsymbol{S}}_{(L,p,\boldsymbol{s})}$ induced

by the parameters vector $a \in \mathbb{R}^s$. Given a fixed point $\boldsymbol{w} \in \mathscr{W}$, for any $a \in \mathbb{R}^s$, we denote $h_{\boldsymbol{w}}(a) = f_a(\boldsymbol{w})$.

For given $L, p \in \mathbb{N}^*$, if $\|\boldsymbol{s}\|_0 = 0$, there is only one function $f \equiv 0$ in $\overline{\boldsymbol{S}}_{(L,p,\boldsymbol{s})}$ so that $V_{(L,p,\boldsymbol{s})} = 0$ by the definition of VC dimension which satisfies the conclusion. Therefore, given $L, p \in \mathbb{N}^*$, we only need to consider the situation where $\|\boldsymbol{s}\|_0 \geq 1$, i.e. $\boldsymbol{s} \in \{0,1\}^{\overline{p}} \backslash \{0\}^{\overline{p}}$.

Given m fixed points $(\boldsymbol{w}_1, t_1), \ldots, (\boldsymbol{w}_m, t_m) \in \mathscr{W} \times \mathbb{R}$, we first study the total number of signs patterns for the ReLU neural network $\overline{\boldsymbol{S}}_{(L,p,\boldsymbol{s})}$ can output when a varies in $\mathbb{R}^s$, i.e.

$$N(m) = \Big| \left\{ (\operatorname{sgn}(h_{\boldsymbol{w}_1}(a) - t_1), \ldots, \operatorname{sgn}(h_{\boldsymbol{w}_m}(a) - t_m)),\ a \in \mathbb{R}^s \right\} \Big|.$$

Once we have knowledge of it, the necessary condition for $V_{(L,p,\boldsymbol{s})}$ being the VC dimension of $\overline{\boldsymbol{S}}_{(L,p,\boldsymbol{s})}$ is to satisfy the inequality

$$2^{V_{(L,p,\boldsymbol{s})}} \leq N\left[V_{(L,p,\boldsymbol{s})}\right],$$

from which we finally deduce the bound for $V_{(L,p,\boldsymbol{s})}$. The idea of bounding $N(m)$ is to construct a partition $\mathcal{S}$ of $\mathbb{R}^s$ such that within each region $S \in \mathcal{S}$, the functions $h_{\boldsymbol{w}_j}(a) - t_j$ $j \in \{1, \ldots, m\}$ are all fixed polynomials of a with a bounded degree.

The partition is constructed layer by layer for each $l \in \{0, \ldots, L\}$ through a sequence of successive refinements $\mathcal{S}_0, \mathcal{S}_1, \ldots, \mathcal{S}_L$ in the following way:

1. $|\mathcal{S}_0| = 1$. For all $l \in \{1, \ldots, L\}$,

$$\begin{cases} |\mathcal{S}_l| = |\mathcal{S}_{l-1}| & , \quad \text{if } \sum_{i=0}^{l-1} s_i = 0, \\ |\mathcal{S}_l| \leq 2\left(\dfrac{2emlp}{\sum_{i=0}^{l-1} s_i}\right)^{\sum_{i=0}^{l-1} s_i} |\mathcal{S}_{l-1}| & , \quad \text{if } \sum_{i=0}^{l-1} s_i \neq 0. \end{cases} \tag{98}$$

2. For each $l \in \{1, \ldots, L\}$ and each $S \in \mathcal{S}_{l-1}$, when a varies in S, the input to each node in response to each $\boldsymbol{w}_j$, $j \in \{1, \ldots, m\}$ in the l-th layer is a fixed polynomial of total degree no more than l in at most $\sum_{i=0}^{l-1} s_i$ variables of a.

We take $\mathcal{S}_0 = \{\mathbb{R}^s\}$. We check that with this choice both two rules mentioned above are satisfied. It is trivial that $|\mathcal{S}_0| = 1$. Moreover, for each fixed $\boldsymbol{w}_j$, $j \in \{1, \ldots, m\}$, the input to each node in the first layer can be written as a linear combination of the parameters in A_0 and b_0. Therefore, it is a fixed polynomial of degree no more than 1 in at most s_0 variables of a. The second rule of constructing the partition is also satisfied. Suppose that we could do such a successive partition up to $l-1$ and get a sequence of refinements $\mathcal{S}_0, \ldots, \mathcal{S}_{l-1}$, we now consider to define $\mathcal{S}_l$, where $1 \leq l \leq L$. For any $\boldsymbol{w}_j$ with $j \in \{1, \ldots, m\}$, $k \in \{1, \ldots, p\}$ and $S \in \mathcal{S}_{l-1}$, we denote $h_{\boldsymbol{w}_j,k,S}(a)$ the input of the k-th node in the l-th layer in response to $\boldsymbol{w}_j$ for

some $a \in S$. By the induction rules, $h_{\boldsymbol{w}_j,k,S}(a)$ is a fixed polynomial of total degree no more than l in at most $\sum_{i=0}^{l-1} s_i$ variables.

If $\sum_{i=0}^{l-1} s_i \neq 0$, for each $S \in \mathcal{S}_{l-1}$, applying Lemma 12 to the collection of polynomials

$$\mathcal{C} = \{h_{\boldsymbol{w}_j,k,S}(a),\ k \in \{1,\ldots,p\},\ j \in \{1,\ldots,m\}\},$$

we know that for $1 \leq l \leq L$, there are at most

$$N_l = 2\left(\frac{2emlp}{\sum_{i=0}^{l-1} s_i}\right)^{\sum_{i=0}^{l-1} s_i}$$

distinct signs patterns when a varies in S. If $\sum_{i=0}^{l-1} s_i = 0$, for any $S \in \mathcal{S}_{l-1}$, any $k \in \{1,\ldots,p\}$ and any $j \in \{1,\ldots,m\}$, $h_{\boldsymbol{w}_j,k,S}(a)$ is zero so that $\mathcal{C}$ only attains one signs pattern and $N_l = 1$. The successive partition is then based on a refinement of $\mathcal{S}_{l-1}$ such that within each region, all the polynomials belonging to $\mathcal{C}$ have fixed signs when a varies. Thus, for each region $S \in \mathcal{S}_{l-1}$, we partition it into at most N_l subregions and get a refined partition $\mathcal{S}_l$ which satisfies the first rule of the partition. To check that $\mathcal{S}_l$ satisfies the second rule, recall that for any $S' \in \mathcal{S}_l$, since the input to any node in the l-th layer is a fixed polynomial in at most $\sum_{i=0}^{l-1} s_i$ variables of degree no more than l and all the polynomials in the collection

$$\{h_{\boldsymbol{w}_j,k,S'}(a),\ k \in \{1,\ldots,p\}, j \in \{1,\ldots,m\}\}$$

have unchanged signs when a varies in S', we have for each $1 \leq l \leq L$, the input to any node in the $(l+1)$-th layer in response to any $\boldsymbol{w}_j$ is a fixed polynomial of degree no more than $l+1$ in at most $\sum_{i=0}^{l} s_i$ variables of a.

We proceed the partition procedure until getting $\mathcal{S}_L$. Since every step of the partition satisfies (98), we derive

$$|\mathcal{S}_L| \leq \prod_{l=1}^{L} N_l. \tag{99}$$

For any $S \in \mathcal{S}_L$, since $s \geq 1$, the output of the whole network in response to any $\boldsymbol{w}_j$ is a fixed polynomial of degree no more than $L+1$ in at most s variables. By Lemma 12 again, we have for any $S \in \mathcal{S}_L$,

$$\begin{aligned} N_{L+1} &= \Big|\{(\operatorname{sgn}(h_{\boldsymbol{w}_1}(a) - t_1), \ldots, \operatorname{sgn}(h_{\boldsymbol{w}_m}(a) - t_m)),\ a \in S\}\Big| \\ &\leq 2\left(\frac{2em(L+1)}{s}\right)^s. \end{aligned} \tag{100}$$

We intersect all these regions with each $S \in \mathcal{S}_L$ which yields a refined partition $\mathcal{S}_{L+1} = \{S_1,\ldots,S_N\}$ over $\mathbb{R}^s$ with $N = \prod_{l=1}^{L+1} N_l$ combining (99) and (100). We denote $p_l = p$ for all $l \in \{1,\ldots,L\}$ and $p_{L+1} = 1$. Let $\bar{l}$ stand for the smallest number belonging to $\{1,\ldots,L+1\}$ such that $\sum_{i=0}^{\bar{l}-1} s_i \geq 1$.

Therefore, for any m arbitrarily chosen $(\boldsymbol{w}_1, t_1), \ldots, (\boldsymbol{w}_m, t_m) \in \mathscr{W} \times \mathbb{R}$,

$$N(m) \leq \sum_{k=1}^{N} \left| \{(\operatorname{sgn}(h_{\boldsymbol{w}_1}(a) - t_1), \ldots, \operatorname{sgn}(h_{\boldsymbol{w}_m}(a) - t_m)),\ a \in S_k\} \right|$$

$$\leq \prod_{l=1}^{L+1} N_l = 2^{L+2-\bar{l}} \left[\prod_{l=\bar{l}}^{L+1} \left(\frac{2emlp_l}{\sum_{i=0}^{l-1} s_i} \right)^{\sum_{i=0}^{l-1} s_i} \right]. \tag{101}$$

For $l \in \{\bar{l}, \ldots, L+1\}$, let us denote

$$c_l = \frac{2emlp_l}{\sum_{i=0}^{l-1} s_i}, \quad \lambda_l = \frac{\sum_{i=0}^{l-1} s_i}{\sum_{l=\bar{l}}^{L+1} \sum_{i=0}^{l-1} s_i}.$$

We then apply Lemma 13 to (101) and obtain

$$\begin{aligned} N(m) &\leq 2^{L+2-\bar{l}} \left(\prod_{l=\bar{l}}^{L+1} c_l^{\lambda_l} \right)^{\sum_{l=\bar{l}}^{L+1} \sum_{i=0}^{l-1} s_i} \leq 2^{L+2-\bar{l}} \left(\sum_{l=\bar{l}}^{L+1} \lambda_l c_l \right)^{\sum_{l=\bar{l}}^{L+1} \sum_{i=0}^{l-1} s_i} \\ &\leq 2^{L+2-\bar{l}} \left(\frac{2em \sum_{l=1}^{L+1} lp_l}{\sum_{l=\bar{l}}^{L+1} \sum_{i=0}^{l-1} s_i} \right)^{\sum_{l=\bar{l}}^{L+1} \sum_{i=0}^{l-1} s_i} \\ &\leq 2^{L+2-\bar{l}} \left(\frac{2em \sum_{l=1}^{L+1} lp_l}{\sum_{l=1}^{L+1} \sum_{i=0}^{l-1} s_i} \right)^{\sum_{l=1}^{L+1} \sum_{i=0}^{l-1} s_i}. \end{aligned}$$

As we have mentioned, by the definition of VC-dimension, it is necessary to have

$$2^{V_{(L,p,\boldsymbol{s})}} \leq N\left[V_{(L,p,\boldsymbol{s})}\right] \leq 2^{L+2-\bar{l}} \left[\frac{2e\left(\sum_{l=1}^{L+1} lp_l\right) V_{(L,p,\boldsymbol{s})}}{\sum_{l=1}^{L+1} \sum_{i=0}^{l-1} s_i} \right]^{\sum_{l=1}^{L+1} \sum_{i=0}^{l-1} s_i}.$$

Provided $L, p \in \mathbb{N}^*$, we have $\sum_{l=1}^{L+1} lp_l \geq 3$ so that $2e(\sum_{l=1}^{L+1} lp_l) \geq 16$. We then apply Lemma 14 with $m = V_{(L,p,\boldsymbol{s})}$, $t = L + 2 - \bar{l}$, $r = 2e(\sum_{l=1}^{L+1} lp_l)$ and $w = \sum_{l=1}^{L+1} \sum_{i=0}^{l-1} s_i$, and obtain

$$\begin{aligned} V_{(L,p,\boldsymbol{s})} &\leq L + 2 - \bar{l} + \left(\sum_{l=1}^{L+1} \sum_{i=0}^{l-1} s_i \right) \log_2 \left[\left(4e \sum_{l=1}^{L+1} lp_l \right) \log_2 \left(2e \sum_{l=1}^{L+1} lp_l \right) \right] \\ &\leq L + \left(\sum_{l=1}^{L+1} \sum_{i=0}^{l-1} s_i \right) \log_2 \left[\left(4e \sum_{l=1}^{L+1} lp_l \right) \log_2 \left(2e \sum_{l=1}^{L+1} lp_l \right) \right] + 1 \\ &\leq (L+1)(s+1) \log_2 \left[2 \left(2e \sum_{l=1}^{L+1} lp_l \right)^2 \right]. \end{aligned}$$

We complete the proof. □

Acknowledgements

The author is grateful to her supervisor Prof. Yannick Baraud for helpful discussions and constructive suggestions.

References

Akakpo, N. (2012). Adaptation to anisotropy and inhomogeneity via dyadic piecewise polynomial selection. *Mathematical Methods of Statistics*, 21(1):1–28.

Antoniadis, A., Besbeas, P. and Sapatinas, T. (2001). Wavelet shrinkage for natural exponential families with cubic variance functions. *Sankhyā. The Indian Journal of Statistics. Series A*, 63(3):309–327.

Baraud, Y. and Birgé, L. (2014). Estimating composite functions by model selection. *Ann. Inst. H. Poincaré Probab. Statist.*, 50(1):285–314.

Baraud, Y. and Birgé, L. (2018). Rho-estimators revisited: General theory and applications. *Ann. Statist.*, 46(6B):3767–3804.

Baraud, Y., Birgé, L., and Sart, M. (2017). A new method for estimation and model selection: ρ-estimation. *Invent. Math.*, 207(2):425–517.

Baraud, Y. and Chen, J. (2020). Robust estimation of a regression function in exponential families. *arXiv*:2011.01657.

Barron, A., Birgé, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413.

Barlett, P. L., Harvey, N., Liaw, C. and Mehrabian, A. (2019). Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17.

Barlett, P. L., Maiorov, V. and Meir, R. (1998). Almost linear VC-dimension bounds for piecewise polynomial networks. *Neural Computation*, 10(8):2159-2173.

Brown, L. D., Cai, T. T., and Zhou, H. H. (2010). Nonparametric regression in exponential families. *Ann. Statist.*, 38(4):2005–2046.

Dahmen, W., DeVore, R. and Scherer, K. (1980). Multidimensional spline approximation. *SIAM J. Numer. Anal.*, 17(3):380–402.

Daubechies, I., DeVore, R., Foucart, S., Hanin, B. and Petrova, G. (2019). Nonlinear approximation and (deep) ReLU networks. *arXiv*:1905.02199.

Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep Learning*. MIT Press, Cambridge.

Hochmuth,R. (2002). Wavelet characterizations for anisotropic Besov spaces. *Appl. Comput. Harmon. Anal.*, 12(2):179–208.

Kolaczyk, E. D. and Nowak, R. D. (2005). Multiscale generalised linear models for nonparametric function estimation. *Biometrika*, 92(1):119–133.

Massart, P. (2007). *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin. Lectures from

the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *Annals of statistics*, 48(4):1875–1897.

Schumaker, L. (1981). *Spline functions: basic theory*. Wiley, New York.

Suzuki, T. and Nitanda, A. (2021). Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic Besov space. In *35th Advances in Neural Information Processing Systems*, NeurIPS.

Triebel, H. (2006). *Theory of Function Spaces III*. Birkhäuser Verlag, Basel.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes. With Applications to Statistics*. Springer Series in Statistics. Springer-Verlag, New York.

Yamaguti, M. and Hata, M. (1983). Weierstrass's function and chaos. *Hokkaido Math. J.*, 12:333-342.

Department of Mathematics,
University of Luxembourg
Maison du nombre
6 avenue de la Fonte
L-4364 Esch-sur-Alzette
Grand Duchy of Luxembourg
Email address: juntong.chen@uni.lu