UNIVERSITÉ DU LUXEMBOURG

PhD-FSTM-2023-004
The Faculty of Science, Technology and Medicine

DISSERTATION

Defence held on 18/01/2023 in Esch-sur-Alzette

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN MATHEMATIQUES

by

Juntong CHEN
Born on 26 March 1992 in Liaoning (China)

ROBUST ESTIMATION IN EXPONENTIAL FAMILIES: FROM THEORY TO PRACTICE

Dissertation defence committee

Dr. Yannick Baraud, dissertation supervisor
Professor, Université du Luxembourg

Dr. Richard Samworth
Professor, University of Cambridge

Dr. Mark Podolskij, Chairman
Professor, Université du Luxembourg

Dr. Nicolas Verzelen
Professor, Université de Montpellier

Dr. Matthieu Lerasle, Vice Chairman
Professor, ENSAE Paris

Abstract

This thesis is a contribution to the topic of estimation in one-parameter exponential families. It contains four chapters where three different estimation strategies have been studied to address this statistical problem. Chapter 1 is an overall introduction to the subject of this dissertation. Each of the later chapter (Chapter 2, 3 and 4) corresponds to a presentation of one of the three strategies for estimation. Chapter 2 is joint work with Yannick Baraud (University of Luxembourg) and is based on the arXiv paper Baraud and Chen (2020). Chapter 3 is based on the arXiv paper Chen (2022) and Chapter 4 is an ongoing work.

In Chapter 1, we present the statistical problem we would like to solve in this thesis. Roughly speaking, we observe n pairs of independent (but not necessarily i.i.d.) random variables $X_1 = (W_1, Y_1), \ldots, X_n = (W_n, Y_n)$ and assume for each $i \in \{1, \ldots, n\}$, the conditional distribution $Q_i^\star(w_i)$ of Y_i given $W_i = w_i$ is not far away from a distribution belonging to some one-parameter exponential family with parameter $\boldsymbol{\gamma}^\star(w_i) \in \mathbb{R}$. Throughout this thesis, our goal is to estimate the n conditional distributions $Q_i^\star(w_i)$. We provide some elementary examples to illustrate the problem we would like to solve and survey the relevant literature. We conclude our contributions together with providing an overview of the contents of each chapter. We also introduce some background knowledge in this part including a brief introduction to ρ-estimation which is the cornerstone of the work in this thesis and the definition of VC-subgraph class which is the main assumption the present work relies on.

In Chapter 2, we present our first strategy, a robust estimation procedure based on one model. Our estimation relies on the assumptions (might not be true) that the data are i.i.d. and the conditional distributions of Y_i given $W_i = w_i$ belong to a one parameter exponential family $\overline{\mathscr{Q}} = \{Q_\theta,\ \theta \in I\}$ with parameter space given by an interval I. More precisely, we pretend that these conditional distributions take the form $Q_{\boldsymbol{\theta}(w_i)} \in \overline{\mathscr{Q}}$ for some $\boldsymbol{\theta}$ belonging to a VC-subgraph class $\overline{\boldsymbol{\Theta}}$ of functions with values in I. For each $i \in \{1, \ldots, n\}$, we estimate $Q_i^\star(w_i)$ by a distribution of the same form, i.e. $Q_{\widehat{\boldsymbol{\theta}}(w_i)} \in \overline{\mathscr{Q}}$, where $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(X_1, \ldots, X_n)$ is a well-chosen estimator with values in $\overline{\boldsymbol{\Theta}}$. We establish non-asymptotic exponential inequalities for the upper deviations of a Hellinger-type distance between the true conditional distributions of the data and the estimated one based on the

exponential family $\overline{\mathscr{Q}}$ and the class of functions $\overline{\boldsymbol{\Theta}}$. We show that our estimation strategy is robust to model misspecification, contamination and the presence of outliers. Besides, when the data are truly i.i.d., the exponential family $\overline{\mathscr{Q}}$ suitably parametrized and the conditional distributions $Q_i^\star(w_i)$ of the form $Q_{\boldsymbol{\theta}^\star(w_i)} \in \overline{\mathscr{Q}}$ for some unknown Hölderian function $\boldsymbol{\theta}^\star$ with values in I, we prove that the estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}^\star$ is minimax (up to a logarithmic factor). Finally, we provide an algorithm for calculating $\widehat{\boldsymbol{\theta}}$ when $\overline{\boldsymbol{\Theta}}$ is a VC-subgraph class of functions of low or moderate dimension and we carry out a simulation study to compare the performance of $\widehat{\boldsymbol{\theta}}$ to that of the MLE and median-based estimators. The proof of our main result relies on an upper bound, with explicit numerical constants, on the expectation of the supremum of an empirical process over a VC-subgraph class. This bound can be of independent interest.

In Chapter 3, we introduce our second estimation strategy that is a model selection procedure based on ρ-estimation. We establish an oracle type inequality for the selected estimator with respect to a Hellinger-type distance. When the data are truly i.i.d., the exponential family $\overline{\mathscr{Q}}$ suitably parametrized and the regression function $\boldsymbol{\gamma}^\star$ exists such that $Q_i^\star(w_i) = R_{\boldsymbol{\gamma}^\star(w_i)}$ for all $i \in \{1, \ldots, n\}$, we show that our estimator $\widehat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}^\star$ is adaptive in the minimax sense over a wide range of anisotropic Besov spaces. In particular, when $\boldsymbol{\gamma}^\star$ has (or is close to) a general additive or multiple index structure, we construct suitable models to approximate this $\boldsymbol{\gamma}^\star$ and prove that the resulted estimators given by our model selection procedure based on these constructed models can circumvent or mitigate the curse of dimensionality. Moreover, we consider the problem of model selection on ReLU neural networks. We provide an example to illustrate that estimating $\boldsymbol{\gamma}^\star$ by our model selection procedure based on neural networks enjoys a much faster converge rate than the one we would obtain by using models based on wavelets. Finally, we apply our model selection procedure to solve variable selection problem in one-parameter exponential families.

When the family of models is very large, the model selection strategy may be extremely costly from a computational point of view. To overcome this difficulty, we consider an alternative strategy in Chapter 4 which is estimator selection. More precisely, we consider the problem of estimator selection in a particular situation where we assume to have an arbitrary collection $\widehat{\boldsymbol{\Gamma}}(\boldsymbol{X}) = \{\widehat{\boldsymbol{\gamma}}_\lambda(\boldsymbol{X}),\ \lambda \in \Lambda\}$ of piecewise constant candidate estimators for the regression function $\boldsymbol{\gamma}^\star$. These estimators are based on our observations $\boldsymbol{X} = (X_1, \ldots, X_n)$ where the dependency of each estimator with respect to the data $\boldsymbol{X}$ may be unknown and we wish to use the same observations namely $\boldsymbol{X}$ to select a suitable estimator denoted as $\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}(\boldsymbol{X})$ among the family $\widehat{\boldsymbol{\Gamma}}(\boldsymbol{X})$. From this point of view, our procedure contrasts with other alternative selection methods based on data splitting, cross validation, hold-out, etc. We establish a non-asymptotic deviation bound that compares the risk of the selected estimator to the infimum of the risks over the collection. We then explain how to apply our procedure to the changepoint detection problem in one-parameter exponential families. The practical performance of our estimator selection procedure is illustrated by

a comparative simulation study under different scenarios and on two real datasets from the copy numbers of DNA and British coal disasters records.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor Yannick Baraud. Without his guidance and encouragement over the past years, it is not possible to have the work in the present thesis. Back in the days in Nice, I was very fortunate to have two statistical courses with him as a master student. Those elegant lectures evoked my strong interest to go further in the field of statistics. However, it is not that easy for me to pursue a PhD at the beginning because of my weak background of mathematical statistics. I still remember when I was in the first year of my PhD, Yannick has been extremely generous with his time and provided numerous support for me. During my study, I have learnt a lot from him not only on mathematical knowledge but also the taste of statistics and precious personalities to become a qualified researcher. I also appreciate a lot for his trust, invaluable advice and careful feedback through the whole PhD training.

It has been a pleasure to be a member of Yannick's research group where it is always full of vibrant atmosphere. I would like to thank my nice colleagues: Alexandre Lecestre, Guillaume Maillard and Hélène Halconruy for their constant encouragement and helpful discussions. I also thank all the people in our department for creating a friendly working environment. My PhD is funded by European Union's Horizon 2020 research and innovation programme under grant agreement N° 811017, which I gratefully acknowledge.

I would like to thank all of my thesis defense committee members. Thanks Prof. Richard Samworth and Prof. Nicolas Verzelen for their careful reading of this thesis. Thanks Prof. Matthieu Lerasle and Prof. Mark Podolskij for a useful discussion. Thanks also to all of my CET members at the University of Luxembourg Prof. Giovanni Peccati and Prof. Ivan Nourdin for their support along this journey.

The past three years is a special period for the whole world. Due to the pandemic, I was not able to go back to my hometown as usual. I thank all my friends in Luxembourg who gave me plenty of warmth and a lot of happiness. In particular, I thank Shi-Yuan Zhou my lovely office mate; Ninghan Chen and Taowen Wang my two big chefs and friends; Xin He and Longfei Song my little sisters and brothers in Belval. I also thank all the members of ULVB with whom I have enjoyed much fun to play volleyball every week.

Finally, I would like to thank my family for their unconditional love and endless support. They are always my intrinsic motivation to become a better person.

Contents

Chapter 1

Introduction

In this dissertation, we consider the following problem. We observe n pairs of independent (but not necessarily i.i.d.) random variables $X_1 = (W_1, Y_1), \dots, X_n = (W_n, Y_n)$ with values in a measurable product space $(\mathscr{W} \times \mathscr{Y}, \mathcal{W} \otimes \mathcal{Y})$. Our aim is to understand how the responses Y_i depend on the covariates W_i. To analyse this dependency, we assume that for each $i \in \{1, \dots, n\}$, the conditional distributions of Y_i given $W_i = w_i$ belong to a one-parameter exponential family $\overline{\mathscr{Q}}$ with parameter $\boldsymbol{\gamma}^\star(w_i) \in \mathbb{R}$. Throughout this thesis, the mapping $\boldsymbol{\gamma}^\star$ will be called the regression function. The most classical example of such a statistical framework is the Gaussian regression one where the variance of the error is known.

Example 1.0.1 (Homoscedastic Gaussian regression with a known variance)**.** The n pairs of independent random variables $(W_1, Y_1), \dots, (W_n, Y_n)$ take their values in $\mathscr{W} \times \mathbb{R}$ (usually $\mathscr{W} \subset \mathbb{R}^d$ with some $d \in \mathbb{N} \backslash \{0\}$),

$$Y_i = \boldsymbol{\gamma}^\star(W_i) + \sigma \varepsilon_i \quad \text{for all } i = 1, \dots, n, \tag{1.0.1}$$

σ is a known positive number, ε_i are unobserved i.i.d. standard real-valued Gaussian random variables and $\boldsymbol{\gamma}^\star : \mathscr{W} \to \mathbb{R}$ is the unknown regression function we want to estimate.

Nevertheless, we want to go beyond this classical Gaussian regression setting and consider more general situations examples of which are given below.

Example 1.0.2 (Binary regression)**.** We observe the clinical characteristics $W_i \in \mathscr{W} \subset \mathbb{R}^d$, $i \in \{1, \dots, n\}$, of n patients and for each of those we report whether or not he/she has developed a given disease D. More precisely, for each $i \in \{1, \dots, n\}$, $Y_i = 1$ if the i-th patient has disease D and $Y_i = 0$ otherwise. Our aim is to estimate the probability of developing the disease knowing the covariate of the patient. To do so, we introduce the following model: the data $(W_1, Y_1), \dots, (W_n, Y_n)$ are i.i.d. and the conditional distribution of Y given $W = w \in \mathscr{W}$ is given by

$$\mathbb{P}\left[Y = y | W = w\right] = \frac{\exp\left[y \boldsymbol{\gamma}^\star(w)\right]}{1 + \exp\left[\boldsymbol{\gamma}^\star(w)\right]} \quad \text{for all } y \in \{0, 1\}, \tag{1.0.2}$$

where the regression function $\boldsymbol{\gamma}^\star$ is assumed to be of the form $\boldsymbol{\gamma}^\star : w \mapsto \langle \eta^\star, w \rangle$ for some unknown vector $\eta^\star \in \mathbb{R}^d$.

Example 1.0.3 (Poisson regression)**.** We want to study the influence of some ecosystems (depending on humidity, temperature, etc.) on the appearance of a rare frog species. For each ecosystem $i \in \{1, \dots, n\}$ with characteristics $W_i \in \mathscr{W} \subset \mathbb{R}^d$, we count the number of frogs $Y_i \in \mathbb{N}$ of that species. We assume the data are i.i.d. and the conditional distribution of Y given $W = w \in \mathscr{W}$ follows a Poisson distribution with mean $\boldsymbol{\gamma}^\star(w) \in (0, +\infty)$, where $\boldsymbol{\gamma}^\star$ belongs to a smoothness class of functions on $\mathscr{W}$ (Hölder, Sobolev, Besov classes for instance).

Example 1.0.4 (Exponential multiplicative regression)**.** We want to study the lifetime of some space equipment as a function of their operating conditions (radiation, temperature, etc.). The random variables W_i are independent taking their values in $\mathscr{W} \subset \mathbb{R}^d$ and for all $i \in \{1, \dots, n\}$

$$Y_i = \frac{Z_i}{\boldsymbol{\gamma}^\star(W_i)},$$

where the Z_i are i.i.d. random variables distributed as an exponential distribution with parameter 1, independently of W_i, and the regression function $\boldsymbol{\gamma}^\star : \mathscr{W} \to (0, +\infty)$ is assumed to be of the form $\boldsymbol{\gamma}^\star : w \mapsto \log(1 + \exp[\langle \eta^\star, w \rangle])$ for some unknown vector $\eta^\star \in \mathbb{R}^d$.

In these examples, we have proposed some particular models for regression functions: a linear set of functions in the case of Example 1.0.2, a parametric set of (nonlinear) functions in the case of Example 1.0.4 and a nonparametric model in the case of Example 1.0.3. Other choices would have been possible.

All these examples can be put in the following general framework: the conditional distributions of Y_i given $W_i = w_i$ are of the form $R_{\boldsymbol{\gamma}^\star(w_i)}$ and belong to a one-parameter exponential family $\overline{\mathscr{Q}} = \{R_\gamma,\ \gamma \in J\}$ where the parameter set J is a non-trivial interval of $\mathbb{R}$ and $\boldsymbol{\gamma}^\star : \mathscr{W} \to J$ belongs to a given class of regression functions $\overline{\boldsymbol{\Gamma}}$. We recall that a one-parameter exponential family with parameter set J is a family of distributions dominated by some reference measure μ and the densities of which take the form for all $y \in \mathscr{Y}$ and $\gamma \in J$

$$\overline{r}_\gamma(y) = e^{u(\gamma)S(y) - B(\gamma)} a(y) \quad \text{where} \quad B(\gamma) = \log\left[\int_{\mathscr{Y}} e^{u(\gamma)S(y)} a(y) d\mu(y)\right], \tag{1.0.3}$$

S 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 J and a is a nonnegative function on $\mathscr{Y}$.

The statistical model can be described as follows: the data (W_i, Y_i) are independent and for each $i \in \{1, \dots, n\}$, (W_i, Y_i) is distributed as $R_{\boldsymbol{\gamma}^\star} \cdot P_{W_i}$ where P_{W_i} denotes the

marginal distribution of W_i and for all measurable sets $A \in \mathcal{Y}$ and $B \in \mathcal{W}$,

$$R_{\boldsymbol{\gamma}^\star} \cdot P_{W_i}(A \times B) = \int_B R_{\boldsymbol{\gamma}^\star(w)}(A) dP_{W_i}(w) = \int_{A \times B} \overline{r}_{\boldsymbol{\gamma}^\star(w)}(y) d\mu(y) dP_{W_i}(w).$$

In the literature, it is often assumed that the W_i are deterministic or that they are i.i.d. with a common distribution P_W.

1.1 The Gaussian regression framework

As already mentioned, the Gaussian case (Example 1.0.1) is probably the most widely studied in the literature. One of the reason lies in the fact, that under suitable assumptions, the maximum likelihood estimator (MLE for short) of the regression function $\boldsymbol{\gamma}^\star$ is both easy to calculate and analyse. This is typically the case when the set $\overline{\boldsymbol{\Gamma}}$ is linear with dimension D and the $W_i = w_i$ are deterministic. Then the MLE is a linear estimator $\widehat{\boldsymbol{\gamma}}_n$ given by the least squares and its quadratic risk satisfies

$$\mathbb{E}\left[d^2(\boldsymbol{\gamma}^\star, \widehat{\boldsymbol{\gamma}}_n)\right] \leq \inf_{\boldsymbol{\gamma} \in \overline{\boldsymbol{\Gamma}}} d^2(\boldsymbol{\gamma}^\star, \boldsymbol{\gamma}) + \frac{D}{n}\sigma^2, \tag{1.1.1}$$

where $d(\boldsymbol{\gamma}, \boldsymbol{\gamma}') = \left[n^{-1}\sum_{i=1}^n (\boldsymbol{\gamma}(w_i) - \boldsymbol{\gamma}'(w_i))^2\right]^{1/2}$. When the regression function $\boldsymbol{\gamma}^\star$ does belong to $\overline{\boldsymbol{\Gamma}}$ (as expected) the bound we get is proportional to its dimension D. However, the risk of the MLE becomes more difficult to analyse when the W_i are random or when the set $\overline{\boldsymbol{\Gamma}}$ is no longer a linear space. When the design is random, the authors (e.g. Barron et al. (1999), Kohler (2000), Baraud (2002) and Schmidt-Hieber (2020)) usually assume that the regression function is bounded by some constant M and the risk bound that they established deteriorates when M is taken as a large number. We are not aware of any non-asymptotic analysis of the MLE for general models $\overline{\boldsymbol{\Gamma}}$ especially parametric ones that are nonlinear.

In what follows, we provide an overview of some problems that have been tackled in the Gaussian setting and that we wish to solve in the more general setting of one-parameter exponential families.

1.1.1 From risk bounds to minimax bounds

It is well-known that inequality (1.1.1) based on the parametric model $\overline{\boldsymbol{\Gamma}}$ can lead to a minimax risk bound on a nonparametric set of regression functions. For example, let $\alpha \in (0, 1]$, $M > 0$ and $\mathcal{H}_\alpha(M)$ denote the set of all functions $\boldsymbol{\gamma}$ on $[0, 1]$ such that

$$|\boldsymbol{\gamma}(x) - \boldsymbol{\gamma}(y)| \leq M|x - y|^\alpha, \quad \text{for all} \quad x, y \in [0, 1]\,.$$

If $\overline{\boldsymbol{\Gamma}}$ is the linear space of piecewise constant functions based on a regular partition of $[0, 1]$ into D intervals and $\boldsymbol{\gamma}^\star \in \mathcal{H}_\alpha(M)$, then

$$\inf_{\boldsymbol{\gamma} \in \overline{\boldsymbol{\Gamma}}} d^2(\boldsymbol{\gamma}^\star, \boldsymbol{\gamma}) \leq M^2 D^{-2\alpha}. \tag{1.1.2}$$

Therefore we deduce from (1.1.1) that when D is the smallest integer such that

$$\left(\frac{nM^2}{\sigma^2}\right)^{\frac{1}{1+2\alpha}} \leq D,$$

the estimator $\widehat{\boldsymbol{\gamma}}_n$ achieves the bound

$$\mathbb{E}\left[d^2(\boldsymbol{\gamma}^\star, \widehat{\boldsymbol{\gamma}}_n)\right] \leq 2\left[\left(\frac{\sigma^2 M^{1/\alpha}}{n}\right)^{\frac{2\alpha}{1+2\alpha}} + \frac{\sigma^2}{n}\right]$$

uniformly over the smoothness class $\mathcal{H}_\alpha(M)$. It turns out this rate is minimax in the sense that it cannot be improved by any estimator uniformly over $\mathcal{H}_\alpha(M)$ apart from the numerical constants. An interesting feature of this strategy lies in the following fact: we have considered a set $\overline{\boldsymbol{\Gamma}}$ of regression functions that is only approximate and may not contain the the true regression function $\boldsymbol{\gamma}^\star$ (but does contain our estimator) in order to get an optimal risk bound on the class of regression functions $\mathcal{H}_\alpha(M)$ of interest. Applying this strategy is possible because inequality (1.1.2) shows that the bound on the quadratic risk of the MLE remains stable under a slight misspecification of the class $\overline{\boldsymbol{\Gamma}}$ of regression functions we have started from. This property allows one to consider models with good approximation properties rather than exact models.

There exist numerous results on how to approximate smooth functions by finite dimensional linear spaces (e.g. based on piecewise polynomials, splines or wavelets see DeVore and Lorentz (1993), Birgé and Massart (1997) and Donoho and Johnstone (1998) for instance). More recently, approximation and estimation by neural networks have received increasing attention in the community of mathematicians in the area of approximation theory and statistics (e.g. Daubechies et al. (2019) and Schmidt-Hieber (2020)).

1.1.2 Structure assumptions

Estimating a regression function under smoothness assumptions as we did in the section above is quite satisfactory when the W_i belong to a subset of $\mathbb{R}^d$ with $d = 1$ but this strategy becomes useless when d is large. It is indeed well-known that the minimax rate over a class of functions of smoothness α on $[0,1]^d$ is typically of order $n^{-2\alpha/(2\alpha+d)}$ (see Stone (1982)). A way to overcome this difficulty is to make structure assumptions on the regression function $\boldsymbol{\gamma}^\star$ namely to assume that the unknown function $\boldsymbol{\gamma}^\star$ is of the form $f \circ g$ where f and g have some specific structures (for instance see Stone (1985), Horowitz and Mammen (2007) and Baraud and Birgé (2014)). One typical example of structure assumptions is the generalized additive structure. More precisely, we may assume that the regression function $\boldsymbol{\gamma}^\star$ on $\mathscr{W} = [0,1]^d$ is of the form $w = (w_1, \ldots, w_d) \mapsto f(g(w_1, \ldots, w_d))$ where f is a smooth function of regularity α on $[0,1]$ and g is of the form

$$g(w_1, \ldots, w_d) = g_1(w_1) + \ldots + g_d(w_d),$$

where the g_i are also of regularity α and take their values in $[0, 1/d]$. In this case, Horowitz and Mammen (2007) shows that, one can estimate the regression function $\boldsymbol{\gamma}^\star$ with rate $n^{-2\alpha/(2\alpha+1)}$ which is independent of the dimension d.

1.1.3 Model selection

Dealing with a single model $\overline{\boldsymbol{\Gamma}}$ for the regression function is usually not enough in most applications. It is preferable to introduce an at most countable family of candidate models $\left\{\overline{\boldsymbol{\Gamma}}_m,\ m \in \mathcal{M}\right\}$ and then to design a criterion solely based on the data $\boldsymbol{X} = (X_1, \ldots, X_n)$ in order to select a suitable index $\widehat{m}(\boldsymbol{X})$ among the family $\mathcal{M}$. In Gaussian regression, Birgé and Massart (2001) proposed a penalized model selection criterion. Their result assumes that the W_i are deterministic and that each $\overline{\boldsymbol{\Gamma}}_m$ is a linear space of finite dimension D_m. The selected index $\widehat{m}(\boldsymbol{X})$ satisfies

$$\mathbb{E}\left[d^2(\boldsymbol{\gamma}^\star, \widehat{\boldsymbol{\gamma}}_{\widehat{m}})\right] \leq C \left\{ \inf_{m \in \mathcal{M}} \left[d^2(\boldsymbol{\gamma}^\star, \overline{\boldsymbol{\Gamma}}_m) + \frac{D_m(L_m \vee 1)}{n}\sigma^2 \right] + \frac{\Sigma}{n}\sigma^2 \right\}, \tag{1.1.3}$$

where C is an explicit numerical constant, $\{L_m\}_{m\in\mathcal{M}}$ is a family of nonnegative numbers satisfying $\Sigma = \sum_{m \in \{m' \in \mathcal{M} | D_{m'} > 0\}} \exp\left[-D_m L_m\right] < +\infty$ and $d(\boldsymbol{\gamma}^\star, \overline{\boldsymbol{\Gamma}}_m) = \inf_{\boldsymbol{\gamma} \in \overline{\boldsymbol{\Gamma}}_m} d(\boldsymbol{\gamma}^\star, \boldsymbol{\gamma})$. With such a result at hand, the authors provide several applications among which variable selection and adaptation in the minimax sense.

We are not aware of any result that generalizes their approach to other exponential families and to possibly nonlinear models $\overline{\boldsymbol{\Gamma}}_m$.

1.1.4 Changepoint detection

The problem of estimating the changepoints of a regression function defined on the bounded interval $\mathscr{W} \subset \mathbb{R}$ can be described as follows: the regression function $\boldsymbol{\gamma}^\star$ is assumed to be piecewise constant on a partition $m^\star$ of $\mathscr{W}$ into a finite number of intervals. The aim is to estimate both the partition $m^\star$ (or equivalently the endpoints of the intervals of $m^\star$) as well as the values of the function $\boldsymbol{\gamma}^\star$ on each interval of the partition. This problem has received a lot attention in the literature. Some authors put more emphasis on the problem of localization of the changepoints (e.g. Fryźlewicz (2014) and Li et al. (2016)) while others consider the problem of estimating the regression function with a small risk (e.g. Baraud et al. (2009) and Cleynen and Lebarbier (2017)). This problem can be viewed as a particular case of model selection where $\mathcal{M}$ is a family of partitions on $[0, 1]$ into a finite number of subintervals and $\overline{\boldsymbol{\Gamma}}_m$ is a linear space of piecewise constant functions based on the partition $m \in \mathcal{M}$.

1.1.5 Estimator selection

When the family of models is very large, which may be the case for solving the variable selection and changepoint detection problems, the model selection strategy proposed by Birgé and Massart (2001) may be extremely costly from a computational point of view. To overcome this difficulty, an alternative approach is to start from a family of preliminary estimators $\{\widehat{\boldsymbol{\gamma}}_\lambda,\ \lambda \in \Lambda\}$, the number of which keeps to a reasonable size, and to design a selection procedure in order to select a suitable one $\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}$ among the family to estimate $\boldsymbol{\gamma}^\star$. These estimators may, for example, result from a more trackable model selection procedure or be obtained based on extra assumptions on the regression function $\boldsymbol{\gamma}^\star$. The problem of selecting a suitable estimator among a collection of candidate ones is called estimator selection. Many procedures tackle this problem on the basis of a sample splitting scheme (e.g. Wegkamp (2003), Yang (2004) and Bunea et al. (2007)). This means that the estimators are built from a sample $\mathcal{S}_A$ and conditionally on $\mathcal{S}_A$, we select one of them by means of an independent sample $\mathcal{S}_B$. Considering the sample $\mathcal{S}_B$ only, the collection of candidate estimators $\{\boldsymbol{\gamma}_\lambda,\ \lambda \in \Lambda\}$ can be seen as non-random. If the regression function $\boldsymbol{\gamma}^\star$ and all the candidates $\boldsymbol{\gamma}_\lambda$ are bounded in sup-norm, Bunea et al. (2007) proved that their selected estimator $\boldsymbol{\gamma}_{\widehat{\lambda}}$ satisfies

$$\mathbb{E}\left[d^2(\boldsymbol{\gamma}^\star, \boldsymbol{\gamma}_{\widehat{\lambda}})\right] \leq (1+\varepsilon) \inf_{\lambda \in \Lambda} d^2(\boldsymbol{\gamma}^\star, \boldsymbol{\gamma}_\lambda) + C_\varepsilon \sigma^2 \frac{\log(\mathrm{Card}(\Lambda))}{n}, \tag{1.1.4}$$

where $C_\varepsilon > 0$ is a constant only depending on ε. The risk of the selected estimator compares to the infimum of the risks among the the family of estimators we started from up to an additional term that increases with the cardinality of the family Λ. They show that this term, namely $\log(\mathrm{Card}(\Lambda))/n$, cannot be removed in general.

Baraud et al. (2014) proposed an alternative approach that relaxes the assumption that the estimators are based on the independent sample $\mathcal{S}_A$, allowing thus both the estimators and the selection rule to be based on the same dataset. Besides, for each specific problem they want to solve (variable selection, selection of a suitable tuning parameter), the risk bound they got on the selected estimator is a non-increasing function of the family Λ (with respect to the inclusion), which means that the risk bound they got can only be improved when one enlarges the family Λ. This might not be the case with a risk bound of the form (1.1.4). To our knowledge, their strategy has never been generalized to other exponential families.

1.2 Regression in other exponential families

We are not aware of many results for estimating the regression function $\boldsymbol{\gamma}^\star$ when the exponential family is not the Gaussian one.

1.2.1 Converting the problem to Gaussian regression

When the W_i are deterministic, a common strategy is to bin the data into groups and to make a suitable transformation of a combination of those into each group so that the distributions of the resulting statistics are close to a Gaussian random variable. This can be obtained by using some variance stabilization transformation (VST) techniques. For instance, Anscombe (1948) proposed VSTs for Poisson, binomial and negative binomial data. After performing VST, one can apply methodologies that have been developed in the Gaussian framework to the transformed data. For instance, when dealing with Poisson data, Donoho (1993) applied the transformation introduced in Anscombe (1948) while Fryźlewicz and Nason (2001, 2004) applied the one given by Fisz (1955). Another example of implementing this strategy can be found in Nunes and Nason (2009) where they dealt with binomial regression. Theoretical guarantee of this strategy mainly relies on asymptotically normal approximations. We refer the reader to Proposition 2 of Fryźlewicz and Nason (2004) and Theorem 3.1 of Nunes and Nason (2009) for instance. It has been reported in Besbeas et al. (2004) that for Poisson data, this strategy may suffer from over-smoothing or losing details of the underlying signals especially when the levels of counts are low.

On a unified treatment of the one-parameter exponential families, Brown et al. (2010) introduced a new transformation technique and established uniform risk bounds on sets of regression functions that belong to Besov spaces. Up to a logarithmic factor, the risk bounds they got coincide with those obtained in the Gaussian setting. However, their approach only applies for exponential families which are parametrized by their means and under the conditions that these means are bounded away from zero and infinity. Furthermore, their results mainly focus on the situation where the variances of the distributions in the exponential family are quadratic functions of their means.

All the above mentioned results are of asymptotic nature. Moreover, a common feature of the techniques that are used in all these papers lies in the fact that they require the W_i to be deterministic in order to bin the data into non-random groups.

1.2.2 Direct treatments on the original non-Gaussian data

In all the literature we have found, the authors developed their procedures for some specific parametrization of the exponential family $\overline{\mathscr{Q}}$. Some of them assume that $\overline{\mathscr{Q}}$ has been parametrized by its mean and they considered the problem of estimating the conditional means $\mathbb{E}[Y_i|W_i = w_i]$, which correspond then to the values of the regression function $\boldsymbol{\gamma}^\star$ at the w_i. Others assume that $\overline{\mathscr{Q}}$ is under its natural form, i.e. taking u in (1.0.3) as the identity function.

To estimate the regression function $\boldsymbol{\gamma}^\star$, some of the authors used wavelet techniques. For instance, Antoniadis and Leblanc (2000) focused on the estimation of the conditional

means $\mathbb{E}[Y_i|W_i = w_i]$ in binary regression by implementing wavelet expansion. They assume that the $W_i = w_i$ are deterministic and the regression function $\boldsymbol{\gamma}^\star$ belongs to a Hölder class of functions of smoothness $\alpha > 1/2$ on $[0,1]$. They showed that their estimator is asymptotically minimax when $\alpha \geq 1$, achieving the rate $n^{-2\alpha/(2\alpha+1)}$ for the squared integrated $\mathbb{L}_2$-loss (with respect to the Lebesgue measure on $[0,1]$). In Ivanoff et al. (2016), the authors considered Poisson regression parametrized by the mean. Besides wavelets, they also considered other classical orthonormal systems such as the Fourier basis. Their estimator of $\mathbb{E}[Y_i|W_i = w_i]$ is based on a penalized likelihood criterion where the penalties are of the form Lasso or group-Lasso ones. Under the assumptions that the true conditional distributions belong to their model and that the conditional means $\mathbb{E}[Y_i|W_i = w_i]$ are bounded away from zero and infinity, they established a risk bound based on the empirical Kullback-Leibler divergence between the true conditional distributions of the data and the estimated one based on $\widehat{\boldsymbol{\gamma}}$. By applying wavelet shrinkage, Antoniadis and Sapatinas (2001) considered the problem of estimating the conditional means $\mathbb{E}[Y_i|W_i = w_i]$ in all the one-parameter exponential families with quadratic variance functions (i.e. the variance of the distribution is at most a quadratic function of its mean). Their approach is inspired by the work of Beran and Dümbgen (1998) on modulation estimators in Gaussian regression. When the regression function $\boldsymbol{\gamma}^\star$ belongs to an ellipsoid of the Sobolev class W_2^α with $\alpha > 1/2$, the estimator proposed by Antoniadis and Sapatinas (2001) attains the classical rate of convergence $n^{-2\alpha/(2\alpha+1)}$ with respect to the squared $\mathbb{L}_2$-loss (with respect to the empirical measure $(\sum_{i=1}^n \delta_{W_i})/n$). Later, Antoniadis et al. (2001) extended the approach in Antoniadis and Sapatinas (2001) to the one-parameter exponential families with cubic variance functions.

Some authors have also considered the problem of variable selection assuming that the regression function $\boldsymbol{\gamma}^\star$ is a sparse linear combination of the covariates of W_i. In order to estimate the subset of covariates that really influence the response Y_i, they used a penalized criterion including the likelihood function and the penalty term which is similar to the Lasso one in the Gaussian setting. This is the case in Li and Cevher (2015) and Jia et al. (2019), where both of the authors parametrized the Poisson distributions under their natural form and showed that their estimators $\widehat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}^\star$ are consistent under some assumptions on the true regression function $\boldsymbol{\gamma}^\star$.

All the results we have mentioned in this section are either established conditionally on the W_i or when the W_i are deterministic. This is not the case of the paper by Kroll (2019) which considered the problem of model selection in Poisson regression. He showed that his estimator of the conditional mean function $\mathbb{E}[Y_i|W_i]$ satisfies an oracle type inequality with respect to the squared $\mathbb{L}_2(P_W)$-loss, where P_W denotes the common distribution of the W_i. Moreover, when the W_i are uniformly distributed on $[0,1]$, he proved that his estimator achieves the minimax rate of convergence adaptively over Sobolev-type ellipsoids. His results are based on the assumptions that the models are finite dimensional linear spaces

satisfying some good connections between the sup-norm and the $\mathbb{L}_2(P_W)$-norm. Besides, the regression function needs to be bounded in sup-norm.

1.3 Our contributions

In this thesis, our aim is to have a unified treatment of the problem of estimating a regression function in the one-parameter exponential families. Besides, we aim at solving this problem under conditions which we wish to be as weak as possible and in particular, under no assumption on the distributions of the covariate W_i. Furthermore, we want to go beyond the common assumption that the true distribution of the data exactly belongs to the statistical model we consider. If we go back to Example 1.0.2, it is actually unclear that all the data are i.i.d. and it may happen that the probability of developing the disease D is different for a small subgroup of people from the rest of the population (some of these people may have a rare mutation on the gene for instance). It may also happen that some of the data have been erroneously reported. This might even be more likely when counting the population of a species in Example 1.0.3 where by mistake an individual is reported twice. This means that not only the model $\overline{\mathbf{\Gamma}}$ for the regression function might not be exact but also the true conditional distributions of Y_i given the covariates W_i might not belong to the exponential family we consider. It is therefore wiser to assume that our statistical model is an approximation of the truth rather than assuming that it perfectly models reality. Unfortunately, the procedures that have been developed in the literature do not consider the situation where the model is possibly misspecified. In order to tackle this problem, our approach is based on the estimation of the conditional distributions of the data rather than on the sole estimation of the regression function. This explains why we work on a slightly different statistical model than the one we introduced at the beginning of this chapter. We denote by $Q_1^\star(w)\dots,Q_n^\star(w)$ the respective true conditional distributions of Y_i given $W_i = w$. On the one hand, each pair (W_i, Y_i) is distributed as $Q_i^\star \cdot P_{W_i}$ for $i = 1,\dots,n$. On the other hand, the statistical model that we consider, and which may only be an approximation of the truth, assumes that these conditional distributions $Q_i^\star(w)$ are of the form $R_{\boldsymbol{\gamma}^\star(w)} \in \overline{\mathscr{Q}}$ for some $\boldsymbol{\gamma}^\star \in \overline{\mathbf{\Gamma}}$.

Within this statistical setting, we first aim at estimating the n-tuple $\mathbf{Q}^\star = (Q_1^\star,\dots,Q_n^\star)$ by an estimator of the form $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}} = (R_{\widehat{\boldsymbol{\gamma}}},\dots,R_{\widehat{\boldsymbol{\gamma}}})$ for some $\widehat{\boldsymbol{\gamma}} \in \overline{\mathbf{\Gamma}}$. We wish to establish an inequality akin to (1.1.1), except for the following facts:

(1) the distance d now measures the distance between two n-tuples $\mathbf{Q} = (Q_1,\dots,Q_n)$ and $\mathbf{Q}' = (Q_1',\dots,Q_n')$ and is defined as

$$d(\mathbf{Q},\mathbf{Q}') = \left[\frac{1}{n}\sum_{i=1}^{n}\int_{\mathscr{W}} h^2(Q_i(w),Q_i'(w))dP_{W_i}(w)\right]^{1/2}, \tag{1.3.1}$$

where $h(\cdot,\cdot)$ denotes the Hellinger distance. We recall that the Hellinger distance between two probabilities $P = p \cdot \lambda$ and $P' = p' \cdot \lambda$ dominated by λ on a measurable space $(E, \mathcal{E})$ is defined as

$$h(P,P') = \left[\frac{1}{2}\int_E (\sqrt{p} - \sqrt{p'})^2 d\lambda\right]^{1/2}. \tag{1.3.2}$$

(2) We want to relax the assumption that $\overline{\mathbf{\Gamma}}$ is a linear space and replace it by the more general one that it is VC-subgraph with dimension not larger than D.

(3) We want to allow the W_i to be either deterministic or random but not necessarily i.i.d.

In order to describe our contributions in a more specific way, we provide below an account of the contents of the Chapter 2, 3 and 4 of this thesis.

1.3.1 An overview of Chapter 2

Chapter 2 is based on joint work with my PhD supervisor Yannick Baraud and has been submitted for publication as Baraud and Chen (2020). In this chapter, we use ρ-estimation to design a suitable estimator $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}} = (R_{\widehat{\boldsymbol{\gamma}}}, \ldots, R_{\widehat{\boldsymbol{\gamma}}})$ of the n-tuple $\mathbf{Q}^\star = (Q_1^\star, \ldots, Q_n^\star)$ on the basis of a model $\overline{\mathbf{\Gamma}}$ for the regression function.

(1) When $\overline{\mathbf{\Gamma}}$ is a VC-subgraph class on $\mathscr{W}$ with dimension bounded by $D \geq 1$, we show that whatever the conditional distributions $\mathbf{Q}^\star = (Q_1^\star, \ldots, Q_n^\star)$ of Y_i given $W_i = w_i$ and the distributions of the W_i, for some universal constant $C > 0$, our estimator satisfies

$$\mathbb{E}\left[d^2(\mathbf{Q}^\star, \mathbf{R}_{\widehat{\boldsymbol{\gamma}}})\right] \leq C\left[\inf_{\boldsymbol{\gamma}\in\overline{\mathbf{\Gamma}}} d^2(\mathbf{Q}^\star, \mathbf{R}_{\boldsymbol{\gamma}}) + \frac{D}{n}\log n\right]. \tag{1.3.3}$$

The quantity $(D/n)\log n$ is the bound we would get if the model were exact that is when for all $i \in \{1,\ldots,n\}$, $Q_i^\star = R_{\boldsymbol{\gamma}^\star}$ for some $\boldsymbol{\gamma}^\star \in \overline{\mathbf{\Gamma}}$. This bound cannot be improved in general. When the model is approximate in the sense that there exists an element $R_{\overline{\boldsymbol{\gamma}}}$ with $\overline{\boldsymbol{\gamma}} \in \overline{\mathbf{\Gamma}}$ such that for all those indices $i \in I \subset \{1,\ldots,n\}$,

$$\int_{\mathscr{W}} h^2(Q_i^\star(w), R_{\overline{\boldsymbol{\gamma}}(w)}) dP_{W_i}(w) \leq \varepsilon^2$$

the risk bound (1.3.3) becomes

$$\mathbb{E}\left[d^2(\mathbf{Q}^\star, \mathbf{R}_{\widehat{\boldsymbol{\gamma}}})\right] \leq C\left[\frac{|I|}{n}\varepsilon^2 + \frac{|I^c|}{n} + \frac{D}{n}\log n\right] \leq C\left[\varepsilon^2 + \frac{|I^c|}{n} + \frac{D}{n}\log n\right],$$

since the Hellinger distance is bounded by 1. If the quantity $\max\{n\varepsilon^2, |I^c|\}$ is small enough compared to $D\log n$, the bound we get is still of order $(D/n)\log n$. This means that if most of the true conditional distributions $Q_i^\star(\cdot)$ are close enough to

a conditional distribution that belongs to our model $\{R_{\boldsymbol{\gamma}},\ \boldsymbol{\gamma} \in \overline{\boldsymbol{\Gamma}}\}$, the risk bound that we get for our estimator is of the same order of magnitude as the one we would get when the statistical model is exact. This property accounts for the stability of our estimator on the model misspecification and we call it robustness.

(2) Interestingly, we also show that the result (1.3.3) we establish is not connected to a special parametrization of the exponential families therefore holds for all parametrizations. In order to be more specific, let us consider the following situation. Two statisticians A and B want to analyse the same data by using the same statistical model except from the fact that the first statistician puts the exponential family under its natural form while the other uses the general form given by (1.0.3). Since the statistical model to analyse the data is the same, if the statistician A considers the set $\overline{\boldsymbol{\Gamma}}$ to model the regression function, statistician B would use the set $\widetilde{\boldsymbol{\Gamma}} = \{u^{-1}(\boldsymbol{\gamma}),\ \boldsymbol{\gamma} \in \overline{\boldsymbol{\Gamma}}\}$. By using our procedure, we will show that both statisticians will end up with the same estimator of the conditional distributions and that the conditions on which our risk bound (1.3.3) holds is invariant with respect to the choice of u, and it is therefore independent on the way that statisticians parametrize the exponential family.

This feature distinguishes our approach from the ones that can be found in the literature and which mainly rely on a smoothness assumption on the regression function. Such an assumption is not independent with respect to the way of which the exponential family is parametrized since the sets of regression functions $\overline{\boldsymbol{\Gamma}}$ and $\widetilde{\boldsymbol{\Gamma}}$ may have different smoothness depending on the choice of u.

(3) When the data are truly i.i.d. and the model is exact, we derive from inequality (1.3.3) a uniform risk bound over the class of α-Hölder regression functions.

We show that under a suitable parametrization, the order of magnitude of the minimax rate is of order $n^{-2\alpha/(1+2\alpha)}$ in all the one-parameter exponential families (Proposition 2.4.3) at least when all the W_i are uniformly distributed on $[0,1]$. Under such parametrizations, we prove our estimators to be minimax, up to a logarithmic factor (Proposition 2.4.2).

We also provide a counterexample to illustrate the fact that without a suitable parametrization of $\overline{\mathscr{Q}}$, the minimax rate of convergence can be different from the typical one $n^{-2\alpha/(1+2\alpha)}$. For a family of Poisson distributions parametrized by their means, the minimax rate over α-Hölder class is of order $n^{-\alpha/(1+\alpha)}$ (Proposition 2.4.4) and our estimator is minimax up to a logarithmic factor (Proposition 2.4.5).

(4) When $\overline{\mathbf{\Gamma}}$ is a VC-subgraph class of functions of low or moderate dimension, we design an algorithm to calculate the ρ-estimator. We carry out a simulation study in the logit, Poisson and exponential regression problems (the source code is available at https://github.com/juntong6/RhoEstimator). We compare its performance to that of the MLE and a median-based one under three different scenarios: when the model is well-specified, when there is an outlier among the observations and when the data are contaminated. If the model is exact, we see that the ρ-estimator recovers the MLE and both estimators perform well. When the model is slightly misspecified either because we add an outlier or because the data are contaminated, the MLE performs very poorly and its risk explodes while the behavior of the ρ-estimator remains stable. The median-based estimator performs poorly as compared to the ρ-estimator when the model is exact and when the dataset contains an outlier. Its performance becomes comparable to that of the ρ-estimator only when the data are contaminated. As compared to the MLE and the median-based one, only the ρ-estimator shows some good and stable estimation properties under these three scenarios.

1.3.2 An overview of Chapter 3

Chapter 3 is based on a slight modification of the arXiv paper Chen (2022).

As already seen, dealing with a single model is not enough. In Chapter 3, we consider the problem of model selection. More precisely, we consider an at most countable family of models $\{\overline{\mathbf{\Gamma}}_m,\ m \in \mathcal{M}\}$ where for each $m \in \mathcal{M}$, $\overline{\mathbf{\Gamma}}_m$ is a VC-subgraph class on $\mathscr{W}$ with dimension not larger than $D_m \geq 1$ and we design a model selection procedure based on ρ-estimation to choose a suitable element $\widehat{\boldsymbol{\gamma}} \in \overline{\mathbf{\Gamma}} = \cup_{m\in\mathcal{M}}\overline{\mathbf{\Gamma}}_m$. Based on each model $\overline{\mathbf{\Gamma}}_m$ for the regression function $\boldsymbol{\gamma}^\star$, we denote the corresponding model for the n-tuple $\mathbf{Q}^\star = (Q_1^\star, \ldots, Q_n^\star)$ as $\overline{\boldsymbol{\mathscr{Q}}}_m = \{\mathbf{R}_{\boldsymbol{\gamma}} = (R_{\boldsymbol{\gamma}}, \ldots, R_{\boldsymbol{\gamma}}),\ \boldsymbol{\gamma} \in \overline{\mathbf{\Gamma}}_m\}$.

(1) We establish an oracle type inequality for the selected estimator $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}}$, which states that no matter what the conditional distributions $\mathbf{Q}^\star = (Q_1^\star, \ldots, Q_n^\star)$ of Y_i given $W_i = w_i$ are and no matter what the distributions of the W_i are, we have

$$\mathbb{E}\left[d^2(\mathbf{Q}^\star, \mathbf{R}_{\widehat{\boldsymbol{\gamma}}})\right] \leq C\left\{\inf_{m\in\mathcal{M}}\left[d^2(\mathbf{Q}^\star, \overline{\boldsymbol{\mathscr{Q}}}_m) + \frac{D_m}{n}\log n + \frac{\Delta(m)}{n}\right] + \frac{\Sigma}{n}\right\} \tag{1.3.4}$$

where $d(\cdot,\cdot)$ is the distance defined by (1.3.1) and $d(\mathbf{Q}^\star, \overline{\boldsymbol{\mathscr{Q}}}_m) = \inf_{\boldsymbol{\gamma}\in\overline{\mathbf{\Gamma}}_m} d(\mathbf{Q}^\star, \mathbf{R}_{\boldsymbol{\gamma}})$, the notation $\Delta(m)$ plays a similar role as the term $L_m D_m$ in (1.1.3), i.e.

$$\Sigma = \sum_{m\in\mathcal{M}} \exp\left[-\Delta(m)\right] < +\infty,$$

and $C > 0$ is a universal constant.

On the one hand, the risk bound (1.3.4) is akin to (1.1.3) obtained from the problem of model selection in Gaussian regression. On the other hand, compared to the result (1.3.3) based on a single model, the inequality (1.3.4) indicates that if for each $m \in \mathcal{M}$, the term $\Delta(m)$ can be well chosen in the sense that it is proportional to the dimension D_m, we are able to select the model achieving the best trade-off between the approximation and the complexity among the collection $\mathcal{M}$.

With the result (1.3.4) at hand, we can provide several applications as we did in the Gaussian case.

(2) We provide applications to the problems of variable selection and adaptation. For the problem of adaptation to anisotropic Besov spaces, we suppose the data are i.i.d., $\mathscr{W} = [0,1]^d$ and $\mathbf{Q}^\star = (R_{\boldsymbol{\gamma}^\star}, \dots, R_{\boldsymbol{\gamma}^\star})$ where $\boldsymbol{\gamma}^\star$ belongs to some Besov space with an unknown anisotropic regularity $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_d) \in (0, +\infty)^d$. We construct our models $\overline{\boldsymbol{\Gamma}}_m$ as the collections of piecewise polynomials on some particular partitions on $[0,1]^d$. Under some suitable parametrizations of $\overline{\mathscr{Q}}$, we show that in all the one-parameter exponential families, our estimator $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}}$ is adaptive in the minimax sense over a wide range of anisotropic Besov spaces and achieves the risk bound, up to a logarithmic factor, of order $n^{-2\overline{\alpha}/(2\overline{\alpha}+d)}$ where $\overline{\alpha}$ is the harmonic mean of $\alpha_1, \dots, \alpha_d$ (Corollary 3.3.1).

(3) In order to overcome the problem of the curse of dimensionality when the covariates take their values in a high dimensional space, we tackle estimation under structure assumptions on the regression functions. We consider two classical models for $\boldsymbol{\gamma}^\star$ which are generalized additive structure and multiple index structure. Under each structure assumption, we construct suitable models $\overline{\boldsymbol{\Gamma}}_m$ to approximate $\boldsymbol{\gamma}^\star$ and implement our model selection procedure to derive an estimator based on the family of constructed models $\{\overline{\boldsymbol{\Gamma}}_m,\ m \in \mathcal{M}\}$. We establish risk bounds for the resulted estimators (Corollary 3.4.1 and 3.4.2). The results state that under a suitable parametrization of the exponential family $\overline{\mathscr{Q}}$, the bounds we get coincide with those derived in the Gaussian regression setting under the same structure assumption. Therefore, by doing so, we circumvent or mitigate the curse of dimensionality.

(4) We also consider the problem of estimating the regression function $\boldsymbol{\gamma}^\star$ by neural networks. To solve this problem, we provide a bound on the VC dimension that depends on the width, the depth and the sparsity of the network. By using suitable networks, we shall see that our estimator may achieve, up to a logarithmic factor, a parametric rate of convergence for estimating some very irregular regression function (that is nowhere differentiable).

1.3.3 An overview of Chapter 4

In Chapter 4, we consider the problem of estimator selection and as it has been explained in Section 1.1.5, we want to design a selection rule for which the selected estimator will satisfy a risk bound which is a non-increasing function (for the inclusion) of the collection of candidate estimators.

We consider this problem in a particular situation where we assume to have an arbitrary collection $\widehat{\mathbf{\Gamma}}(\boldsymbol{X}) = \{\widehat{\boldsymbol{\gamma}}_\lambda(\boldsymbol{X}),\ \lambda \in \Lambda\}$ of piecewise constant candidates for the regression function $\boldsymbol{\gamma}^\star$ based on the observations $\boldsymbol{X}$. With the same data $\boldsymbol{X}$, our goal is to select an estimator $\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}(\boldsymbol{X})$ among the collection $\widehat{\mathbf{\Gamma}}(\boldsymbol{X})$. Our method is agnostic to the dependencies of each $\widehat{\boldsymbol{\gamma}}_\lambda(\boldsymbol{X})$ with respect to the data $\boldsymbol{X}$ and can therefore be unknown. From this point of view, our procedure contrasts with other alternative selection methods based on data splitting, cross validation, hold-out, etc. but can be regarded as a generalization of the approach in Baraud et al. (2014) from the problem of Gaussian estimator selection to a unified treatment of the problem of estimator selection in all one-parameter exponential families.

(1) Under some reasonable assumptions (Assumption 4.2.1 and 4.2.2), we show the selected estimator $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}}$ of the n-tuple $\mathbf{Q}^\star = (Q_1^\star, \ldots, Q_n^\star)$ satisfies

$$\mathbb{E}\left[d^2(\mathbf{Q}^\star, \mathbf{R}_{\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}})\right] \leq C \inf_{\lambda\in\Lambda}\left\{\mathbb{E}\left[d^2(\mathbf{Q}^\star, \mathbf{R}_{\widehat{\boldsymbol{\gamma}}_\lambda})\right] + \frac{1}{n}\mathbb{E}\left[\mathbf{\Xi}(\widehat{\boldsymbol{\gamma}}_\lambda)\right]\right\}, \tag{1.3.5}$$

where $C > 0$ is a numerical constant only depending on some parameters required in Assumption 4.2.1 and 4.2.2, $\mathbf{\Xi}(\widehat{\boldsymbol{\gamma}}_\lambda)$ is an additional nonnegative term mainly related to the VC dimension of the functional space to which $\widehat{\boldsymbol{\gamma}}_\lambda$ belongs (see Corollary 4.2.1 for details).

Similarly to the risk bound for the selected estimator in Baraud et al. (2014), the result (1.3.5) compares the risk of the selected estimator $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}}$ to those of $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}_\lambda}$ plus an additional nonnegative term which does not depend on the cardinality of the set $\widehat{\mathbf{\Gamma}}$. Therefore, if we enlarge the family of candidates $\widehat{\mathbf{\Gamma}}$, the risk bound for the selected estimator can only be improved.

(2) We then apply our procedure to solve the problem of changepoint detection. To do so, we first calibrate some tuning parameter in our estimator selection procedure based on the simulation study. Our calibration differs from the typical procedures which choose the tuning parameter by cross-validation and have to be done for each implementation. In fact, the parameter we would like to calibrate is a universal constant and we can even derive a possible value of it from our theory. Unfortunately, this theoretical value is too large to use in practice and we do not have enough information about the smallest possible value of it validating (1.3.5). Therefore, we

regard it as a tuning parameter to be calibrated once for all.

(3) We solve the problem of Gaussian changepoint detection with respect to the means by "standing upon the shoulders of giants". To be more precise, we construct our candidates set $\widehat{\mathbf{\Gamma}}(\boldsymbol{X})$ as a collection of several state-of-art procedures in the literature and implement our selection procedure to pick one among $\widehat{\mathbf{\Gamma}}(\boldsymbol{X})$ based on the observations $\boldsymbol{X}$. In Section 4.4, we test the performance of our selection procedure under six different formats of signals. It turns out that under different test signals, our estimator selection procedure tends to allocate different preference to the candidates in $\widehat{\mathbf{\Gamma}}(\boldsymbol{X})$ based on their practical performance. As a result, our estimator provides a very competitive performance under all the six signals in the meanwhile no single procedure (state-of-art estimator in the literature) in $\widehat{\mathbf{\Gamma}}(\boldsymbol{X})$ succeeds to achieve this. Moreover, when there are a small amount of outliers in the observations, we still can select the most competitive ones which implies a robustness property of our selection procedure.

(4) We also consider the application of our procedure to the changepoint detection problem in other exponential families, where the algorithms are much less as compared to the Gaussian case. We are only ware of Cleynen and Lebarbier (2014, 2017) and Frick et al. (2013) which try to solve the changepoint detection problem in general exponential families. To construct a rich collection $\widehat{\mathbf{\Gamma}}$ as we did in the Gaussian changepoint detection, we implement the mean-matching variance stabilizing transformation proposed in Brown et al. (2010) to roughly turn the problem into the Gaussian case and borrow those previous algorithms to locate changepoints. To enhance the robustness with respect to outliers, we associate the ρ-estimator on each resulted partition given by those algorithms. Simulation results in Section 4.4.3 indicate that by doing so, when there is no outlier, our estimator performs much better than the one given in Frick et al. (2013) and slightly outperforms the one given in Cleynen and Lebarbier (2014, 2017). Moreover, when there is a small amount of outliers in the observations, we obviously improve the stability of the final estimator as compared to those two existing methods.

(5) At the end of Chapter 4, we apply our selection procedure to two real datasets including the copy numbers of DNA and British coal disasters records to investigate its practical performance. Gaussian model is considered to detect changes for the first dataset and Poisson model is applied to the second one. For both of them, our estimator shows a reasonable performance according to the relevant literature.

1.4 A brief introduction to ρ-estimation

As mentioned, we shall handle the estimation problem introduced in Section 1.3 by a new methodology which is called ρ-estimation proposed by Baraud et al. (2017). In Baraud and Birgé (2018), the authors revisited this estimation method and relaxed several limitations in the previous work.

This new estimation methodology is designed for providing a universal treatment of a rather general estimation framework. More precisely, they consider the problem that they observe n independent random variables $X_1, \ldots, X_n$ with values in a measurable space $(\mathscr{X}, \mathcal{X})$ where for each $i \in \{1, \ldots, n\}$, X_i follows an unknown distribution $P_i^\star$ on $(\mathscr{X}, \mathcal{X})$. Their goal is to find a suitable random approximation $\widehat{\mathbf{P}}(\boldsymbol{X}) = \otimes_{i=1}^n \widehat{P}_i(\boldsymbol{X})$ of the true joint distribution $\mathbf{P}^\star = \otimes_{i=1}^n P_i^\star$ on the basis of the observations $\boldsymbol{X} = (X_1, \ldots, X_n)$.

1.4.1 Robust and optimal estimators

To motivate the employment of the ρ-estimator, we introduce two essential properties of it: robustness and (nearly) optimality.

We denote $\mathscr{P}^f$ the set of all product probabilities on $(\mathscr{X}^n, \mathcal{X}^{\otimes n})$. To measure the deviation of $\widehat{\mathbf{P}}(\boldsymbol{X})$ from the truth $\mathbf{P}^\star$, a distance on $\mathscr{P}^f$ is needed. A convenient choice of measuring the distance between two product probabilities could be the Hellinger-type distance $\mathbf{h}$ (e.g. Le Cam (1986) and Le Cam and Yang (1990)) defined as for any $\mathbf{P} = \otimes_{i=1}^n P_i \in \mathscr{P}^f$ and $\mathbf{P}' = \otimes_{i=1}^n P_i' \in \mathscr{P}^f$,

$$\mathbf{h}^2\left(\mathbf{P}, \mathbf{P}'\right) = \sum_{i=1}^n h^2(P_i, P_i'), \tag{1.4.1}$$

where $h(\cdot, \cdot)$ denotes the Hellinger distance defined by (1.3.2). We then can quantify the performance of an estimator $\widehat{\mathbf{P}}(\boldsymbol{X})$ through its risk $\mathbb{E}\left[\mathbf{h}^2\left(\mathbf{P}^\star, \widehat{\mathbf{P}}(\boldsymbol{X})\right)\right]$ where the expectation is taken under the true joint distribution $\mathbf{P}^\star$. Let us remark here that for any $\widehat{\mathbf{P}}(\boldsymbol{X}) \in \mathscr{P}^f$, its risk at any $\mathbf{P}^\star \in \mathscr{P}^f$ is naturally bounded by n.

To estimate $\mathbf{P}^\star$, ρ-estimation suggests to work based on the models. For each model $\overline{\mathscr{P}}$, they mean a moderate subset of $\mathscr{P}^f$ for the existence of an estimator $\widehat{\mathbf{P}}$ such that $\sup_{\mathbf{P}^\star \in \overline{\mathscr{P}}} \mathbb{E}\left[\mathbf{h}^2(\mathbf{P}^\star, \widehat{\mathbf{P}})\right]$ is substantially smaller than n. Although they do not assume $\mathbf{P}^\star \in \overline{\mathscr{P}}$ is true, by constructing a model $\overline{\mathscr{P}}$ for $\mathbf{P}^\star$, they do as if $\mathbf{P}^\star$ did belong to $\overline{\mathscr{P}}$ and derive their estimators of $\mathbf{P}^\star$ within $\overline{\mathscr{P}}$. We consider two possible situations below.

Optimality in the minimax sense. When we do have $\mathbf{P}^\star \in \overline{\mathscr{P}}$, a nice criterion of evaluating the performance of $\widehat{\mathbf{P}}$, as mentioned, is to consider its maximal quadratic risk $\sup_{\mathbf{P}^\star \in \overline{\mathscr{P}}} \mathbb{E}\left[\mathbf{h}^2(\mathbf{P}^\star, \widehat{\mathbf{P}})\right]$. We would like to compare it with the minimax one over $\overline{\mathscr{P}}$ defined as $R\left(\overline{\mathscr{P}}\right) = \inf_{\widehat{\mathbf{P}}} \sup_{\mathbf{P}^\star \in \overline{\mathscr{P}}} \mathbb{E}\left[\mathbf{h}^2(\mathbf{P}^\star, \widehat{\mathbf{P}})\right]$, where the infimum runs over all the

possible estimators $\widehat{\mathbf{P}}$. We say an estimator $\widehat{\mathbf{P}}$ is approximately optimal over $\overline{\mathscr{P}}$ in the minimax sense if

$$\sup_{\mathbf{P}^\star \in \overline{\mathscr{P}}} \mathbb{E}\left[\mathbf{h}^2(\mathbf{P}^\star, \widehat{\mathbf{P}})\right] \leq C R\left(\overline{\mathscr{P}}\right), \tag{1.4.2}$$

where C is a constant (ideally being not large, independent of n and the model $\overline{\mathscr{P}}$).

Robustness. In practice, we cannot check precisely whether $\mathbf{P}^\star \in \overline{\mathscr{P}}$ or not. When there is a slight misspecification, i.e. $\mathbf{P}^\star \notin \overline{\mathscr{P}}$ but $\inf_{\mathbf{P}\in\overline{\mathscr{P}}} \mathbf{h}^2\left(\mathbf{P}^\star, \mathbf{P}\right)$ being small, if the bound (1.4.2) remains approximately true, we say such an estimator $\widehat{\mathbf{P}}$ possesses robustness.

Putting these two ingredients together, an estimator $\widehat{\mathbf{P}}$ based on the model $\overline{\mathscr{P}}$ is said to be approximately optimal and robust if for all $\mathbf{P}^\star \in \mathscr{P}^f$,

$$\mathbb{E}\left[\mathbf{h}^2(\mathbf{P}^\star, \widehat{\mathbf{P}})\right] \leq C\left[\inf_{\mathbf{P}\in\overline{\mathscr{P}}} \mathbf{h}^2\left(\mathbf{P}^\star, \mathbf{P}\right) \vee R\left(\overline{\mathscr{P}}\right)\right], \tag{1.4.3}$$

where C is a universal constant.

1.4.2 Heuristic ideas

In this section, we introduce some heuristic ideas behind ρ-estimation in our setting. For a more detailed explanation, we refer to Baraud et al. (2017)[Section 1.4 and Section 3.1 (density framework)] and Baraud and Birgé (2018)[Section 2.5].

The invention of ρ-estimation is based on a development of T-estimation proposed by Birgé (2006), of which the main idea is to first discretise the model $\overline{\mathscr{P}}$ at some scale (with respect to $\mathbf{h}$) then to design robust tests between the balls centred at these discretised points. Alternatively, ρ-estimation suggests to construct tests indicating that, given any two choices in the model (without discretisation), which one is closer to the truth with respect to the distance $\mathbf{h}$. Several assumptions required by T-estimation can therefore be relaxed under this new construction.

Recall that in our setting for each pair $X_i = (W_i, Y_i)$, it follows the distribution $P_i^\star = Q_i^\star \cdot P_{W_i}$ on $(\mathscr{X}, \mathcal{X}) = (\mathscr{W}\times\mathscr{Y}, \mathcal{W}\otimes\mathcal{Y})$. We consider $\overline{\mathbf{\Gamma}}$ a collection of measurable functions from $\mathscr{W}$ into J which gives a model $\overline{\mathscr{P}} = \left\{\mathbf{P}_{\boldsymbol{\gamma}},\ \boldsymbol{\gamma}\in\overline{\mathbf{\Gamma}}\right\} = \left\{\otimes_{i=1}^n \left(\overline{r}_{\boldsymbol{\gamma}}\cdot\mu\cdot P_{W_i}\right),\ \boldsymbol{\gamma}\in\overline{\mathbf{\Gamma}}\right\}$ with $\overline{r}_{\boldsymbol{\gamma}}$ given by (1.0.3) for the true joint distribution $\mathbf{P}^\star = \otimes_{i=1}^n P_i^\star$.

Provided two choices $\boldsymbol{\gamma}, \boldsymbol{\gamma}' \in \overline{\mathbf{\Gamma}}$ which induce two probabilities $P_{i,\boldsymbol{\gamma}} = R_{\boldsymbol{\gamma}}\cdot P_{W_i}$ and $P_{i,\boldsymbol{\gamma}'} = R_{\boldsymbol{\gamma}'}\cdot P_{W_i}$ on $\mathscr{X} = \mathscr{W}\times\mathscr{Y}$, to know which one is closer to $P_i^\star$, a natural consideration is to construct an estimator of the quantity $h^2(P_i^\star, P_{i,\boldsymbol{\gamma}}) - h^2(P_i^\star, P_{i,\boldsymbol{\gamma}'})$. The interesting point lies in the fact that if one can design a "nice" statistic say $T(X_i, \boldsymbol{\gamma}, \boldsymbol{\gamma}')$ to estimate this difference, we can go further than just telling the preference between $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$. In fact, if $T(X_i, \boldsymbol{\gamma}, \boldsymbol{\gamma}')$ is a "good" estimator of $h^2(P_i^\star, P_{i,\boldsymbol{\gamma}}) - h^2(P_i^\star, P_{i,\boldsymbol{\gamma}'})$, taking the supremum with respect to $\boldsymbol{\gamma}'$ over $\overline{\mathbf{\Gamma}}$, we note that

$$\sup_{\boldsymbol{\gamma}'\in\overline{\mathbf{\Gamma}}} T(X_i, \boldsymbol{\gamma}, \boldsymbol{\gamma}') \approx \sup_{\boldsymbol{\gamma}'\in\overline{\mathbf{\Gamma}}} \left[h^2(P_i^\star, P_{i,\boldsymbol{\gamma}}) - h^2(P_i^\star, P_{i,\boldsymbol{\gamma}'})\right] = h^2(P_i^\star, P_{i,\boldsymbol{\gamma}}) - \inf_{\boldsymbol{\gamma}'\in\overline{\mathbf{\Gamma}}} h^2(P_i^\star, P_{i,\boldsymbol{\gamma}'}),$$

which measures how far away for $P_{i,\boldsymbol{\gamma}}$ being the closest element to $P_i^\star$ among $\overline{\mathscr{P}}$. Therefore, to search the closest point to $P_i^\star$ within $\overline{\mathscr{P}}$, the idea is try to, more or less, minimize the quantity $\sup_{\boldsymbol{\gamma}'\in\overline{\boldsymbol{\Gamma}}} T(X_i,\boldsymbol{\gamma},\boldsymbol{\gamma}')$ with respect to $\boldsymbol{\gamma}$. This is the underlying idea to construct the ρ-estimator. At this moment, the only problem left is to design a "good" statistic $T(X_i,\boldsymbol{\gamma},\boldsymbol{\gamma}')$.

To control the risk of the resulted ρ-estimator, the authors suggest to construct $T(X_i,\boldsymbol{\gamma},\boldsymbol{\gamma}')$ as

$$T(X_i,\boldsymbol{\gamma},\boldsymbol{\gamma}') = \psi\left(\sqrt{\frac{\overline{r}_{\boldsymbol{\gamma}'(W_i)}(Y_i)}{\overline{r}_{\boldsymbol{\gamma}(W_i)}(Y_i)}}\right),$$

where two specific choices of the function ψ are

$$\psi_1(x) = \frac{x-1}{\sqrt{x^2+1}} \quad \text{and} \quad \psi_2(x) = \frac{x-1}{x+1}$$

mapping $[0,+\infty]$ into $[-1,1]$. For both of them, they showed that for each $i\in\{1,\ldots,n\}$, whatever $\mathbf{P}^\star \in \mathscr{P}^f$

$$\begin{aligned} a_1 h^2(P_i^\star, P_{i,\boldsymbol{\gamma}}) - a_0 h^2(P_i^\star, P_{i,\boldsymbol{\gamma}'}) &\le \mathbb{E}\left[T(X_i,\boldsymbol{\gamma},\boldsymbol{\gamma}')\right] \\ &\le a_0 h^2(P_i^\star, P_{i,\boldsymbol{\gamma}}) - a_1 h^2(P_i^\star, P_{i,\boldsymbol{\gamma}'}), \end{aligned} \tag{1.4.4}$$

where for ψ_1, $a_0 = 4.97$, $a_1 = 0.083$ and for ψ_2, $a_0 = 4$, $a_1 = 3/8$.

Therefore, given n observations $\boldsymbol{X} = (X_1,\ldots,X_n)$, if we design the statistic $\mathbf{T}(\boldsymbol{X},\boldsymbol{\gamma},\boldsymbol{\gamma}')$ on $(\mathscr{X}^n,\overline{\boldsymbol{\Gamma}},\overline{\boldsymbol{\Gamma}})$ as

$$\mathbf{T}(\boldsymbol{X},\boldsymbol{\gamma},\boldsymbol{\gamma}') = \sum_{i=1}^n T(X_i,\boldsymbol{\gamma},\boldsymbol{\gamma}') = \sum_{i=1}^n \psi\left(\sqrt{\frac{\overline{r}_{\boldsymbol{\gamma}'(W_i)}(Y_i)}{\overline{r}_{\boldsymbol{\gamma}(W_i)}(Y_i)}}\right),$$

an immediate consequence of (1.4.1) and (1.4.4) is that

$$a_1\mathbf{h}^2(\mathbf{P}^\star,\mathbf{P}_{\boldsymbol{\gamma}}) - a_0 h^2(\mathbf{P}^\star,\mathbf{P}_{\boldsymbol{\gamma}'}) \le \mathbb{E}\left[\mathbf{T}(\boldsymbol{X},\boldsymbol{\gamma},\boldsymbol{\gamma}')\right] \le a_0\mathbf{h}^2(\mathbf{P}^\star,\mathbf{P}_{\boldsymbol{\gamma}}) - a_1\mathbf{h}^2(\mathbf{P}^\star,\mathbf{P}_{\boldsymbol{\gamma}'}). \tag{1.4.5}$$

Such a result indicates if the difference between $\mathbf{h}^2(\mathbf{P}^\star,\mathbf{P}_{\boldsymbol{\gamma}})$ and $\mathbf{h}^2(\mathbf{P}^\star,\mathbf{P}_{\boldsymbol{\gamma}'})$ is obviously large, the sign of $\mathbb{E}\left[\mathbf{T}(\boldsymbol{X},\boldsymbol{\gamma},\boldsymbol{\gamma}')\right]$ contains the information about the fact that between $\mathbf{P}_{\boldsymbol{\gamma}}$ and $\mathbf{P}_{\boldsymbol{\gamma}'}$, which one is closer to the truth $\mathbf{P}^\star$.

1.5 VC-subgraph and its dimension

In this section, we introduce some background knowledge of the VC-subgraph including its definition and some useful properties to make a preparation for the contents of Chapter 2, 3 and 4.

Let $\mathcal{C}$ be a collection of subsets of a set $\mathscr{X}$. For an arbitrary set of m points $\{x_1,\ldots,x_m\}$, we say the set $\mathcal{C}$ *shatters* $\{x_1,\ldots,x_m\}$ if each of its 2^m subsets can be represented as a set of the form $C\cap\{x_1,\ldots,x_m\}$ with some $C\in\mathcal{C}$. We begin with introducing the VC-class of sets which was first studied by Vapnik and Chervonenkis.

Definition 1.5.1 (van der Vaart and Wellner (1996), page 134–135)**.** Let $\mathcal{C} = \{C,\ C \subset \mathscr{X}\}$ be a class of subsets of $\mathscr{X}$. We say the VC dimension of $\mathcal{C}$ is $V(\mathcal{C})$ if $V(\mathcal{C})$ is the largest cardinality of the set $\mathcal{E} \subset \mathscr{X}$ which can be shattered by $\mathcal{C}$. Moreover, when $V(\mathcal{C}) < +\infty$, we call $\mathcal{C}$ a VC-class of sets.

To illustrate Definition 1.5.1, let us consider an easy example. Supposing $\mathscr{X} = \mathbb{R}$ and $\mathcal{C}_- = \{(-\infty, a],\ a \in \mathbb{R}\}$, we observe that $\mathcal{C}_-$ shatters any single-point set $\{x\} \subset \mathbb{R}$ but for any two-point set $\{x_1, x_2\} \subset \mathbb{R}$ with $x_1 < x_2$, there is no element $C \in \mathcal{C}_-$ such that $\{x_2\}$ can be represented as the form $C \cap \{x_1, x_2\}$. Therefore, we conclude $V(\mathcal{C}_-) = 1$. Let $\mathcal{C}_+ = \{(b, +\infty),\ b \in \mathbb{R}\}$ and $\mathcal{C} = \mathcal{C}_- \cup \mathcal{C}_+$. Similarly, we can conclude $V(\mathcal{C}) = 2$. This simple example tells that the more refined $\mathcal{C}$ is, the larger is its VC dimension.

We then introduce the conception of growth function $\Pi_{\mathcal{C}}(m)$ of $\mathcal{C}$ defined as

$$\Pi_{\mathcal{C}}(m) = \max_{\{x_1,\ldots,x_m\} \subset \mathscr{X}} \operatorname{Card}\left(\{C \cap \{x_1, \ldots, x_m\},\ C \in \mathcal{C}\}\right). \tag{1.5.1}$$

Based on (1.5.1), the VC dimension of $\mathcal{C}$ can be defined more formally through

$$V(\mathcal{C}) = \sup\left\{m \text{ such that } \Pi_{\mathcal{C}}(m) = 2^m\right\},$$

which turns out to be a convenient tool when one wants to derive an upper bound on the dimension of some class $\mathcal{C}$.

We now move to the definition of VC-subgraph class. Recall that for a function $f : \mathscr{X} \to \mathbb{R}$, its subgraph is a subset of $\mathscr{X} \times \mathbb{R}$ which can be written as

$$\{(x, t) \in \mathscr{X} \times \mathbb{R} \text{ such that } f(x) > t\}.$$

Definition 1.5.2 (van der Vaart and Wellner (1996), page 141)**.** A collection $\mathcal{F}$ of measurable functions on a sample space is called a VC-subgraph class or VC-class if $\mathcal{C}_{\mathcal{F}}$ the collection of all subgraphs of the functions in $\mathcal{F}$ forms a VC-class of sets. Moreover, we say $\mathcal{F}$ is VC-subgraph class with dimension $V(\mathcal{F}) = m$ if $V(\mathcal{C}_{\mathcal{F}}) = m$.

The following result relates the VC-property to any finite-dimensional vector space.

Proposition 1.5.1 (van der Vaart and Wellner (1996), Lemma 2.6.15)**.** *Any finite-dimensional vector space $\mathcal{F}$ of measurable functions $f : \mathscr{X} \to \mathbb{R}$ with dimension $d(\mathcal{F})$ is VC-subgraph of dimension smaller than or equal to $d(\mathcal{F}) + 1$.*

We also list some important properties as follows which allow us to construct new VC-subgraph classes based on those classical ones and establish a dimensional bound for them.

Proposition 1.5.2 (Baraud et al. (2017), Proposition 42)**.** *Let $\mathcal{F}$ be VC-subgraph with dimension V on a set $\mathscr{X}$.*

(i) *For all functions g on $\mathscr{X}$, $\mathcal{F}+g=\{f+g,\ f\in\mathcal{F}\}$ is VC-subgraph with dimension not larger than V.*

(ii) *For all monotone function φ on $\mathbb{R}$, $\varphi(\mathcal{F})=\{\varphi\circ f,\ f\in\mathcal{F}\}$ is VC-subgraph with dimension not larger than V.*

(iii) *The class $-\mathcal{F}$ is VC-subgraph with dimension not larger than V.*

(iv) *The class $\mathcal{F}_+=\{f\vee 0,\ f\in\mathcal{F}\}$ is VC-subgraph with dimension not larger than V.*

We comment that as a consequence of (i) and (iv) of Proposition 1.5.2, under the condition that $V(\mathcal{F})=V$, for any fixed number $a\in\mathbb{R}$, the class $\mathcal{F}_{a+}=\{f\vee a,\ f\in\mathcal{F}\}$ is a VC-subgraph class with dimension not larger than V. Moreover, based on this comment and (iii), for any fixed number $b\in\mathbb{R}$, the class $\mathcal{F}_{b-}=\{f\wedge b,\ f\in\mathcal{F}\}$ is also a VC-subgraph class with dimension not larger than V.

Next theorem states the stability of VC-property under some particular operations which we shall repeatedly use in this thesis.

Theorem 1.5.1 (van der Vaart and Wellner (2009), Theorem 1.1)**.** *Suppose that $\mathcal{C}_1,\dots,\mathcal{C}_m$ are VC-classes of subsets of a given set $\mathscr{X}$ with dimensions $V_1,\dots,V_m$ respectively. We define the classes $\sqcup_{j=1}^m\mathcal{C}_j$ and $\sqcap_{j=1}^m\mathcal{C}_j$ by*

$$\sqcup_{j=1}^m\mathcal{C}_j=\left\{\cup_{j=1}^m C_j,\ C_j\in\mathcal{C}_j,\ j=1,\dots,m\right\},$$

$$\sqcap_{j=1}^m\mathcal{C}_j=\left\{\cap_{j=1}^m C_j,\ C_j\in\mathcal{C}_j,\ j=1,\dots,m\right\}.$$

Then $\sqcup_{j=1}^m\mathcal{C}_j$ and $\sqcap_{j=1}^m\mathcal{C}_j$ are again VC-classes with the following dimensional bounds:

$$\left\{\begin{matrix}V(\sqcup_{j=1}^m\mathcal{C}_j)\\ V(\sqcap_{j=1}^m\mathcal{C}_j)\end{matrix}\right\}\le c_1V\log(c_2m),$$

where $c_1=e/\left[(e-1)\log 2\right]$, $c_2=e/(\log 2)$ and $V=\sum_{j=1}^m V_j$.

Chapter 2

Robust estimation based on a single model

2.1 Introduction

We start this chapter with a more straightforward example to tell the drawbacks of implementing MLE in practice.

Example 2.1.1 (Logit regression)**.** We study a cohort of n patients with respective clinical characteristics $W_1, \dots, W_n$ with values in $\mathbb{R}^d$. For the sake of simplicity we shall assume that d is small compared to n even though this situation might not be the practical one. We associate the label $Y_i = 1$ to the patient i if she/he develops the disease D and $Y_i = -1$ otherwise. The effect of the clinical characteristic W on the probability of developing the disease D is given by the conditional distribution of Y given W: $\mathbb{P}[Y = y|W = w]$ which is the quantity we want to estimate. A classical model for it is the logit one given by

$$\mathbb{P}[Y = y|W = w] = \frac{1}{1 + \exp[-y\langle w^\star, w\rangle]} \in (0,1) \quad \text{for } y \in \{-1, +1\}, \tag{2.1.1}$$

where $w^\star$ is an unknown vector and $\langle \cdot, \cdot \rangle$ the inner product of $\mathbb{R}^d$. If we assume that this model is true, the problem amounts to estimate $w^\star$ on the basis of the observations (W_i, Y_i) for $i \in \{1, \dots, n\}$.

A common way of solving this problem is to use the MLE. In exponential families, the MLE is known to enjoy many nice properties but it also suffers from several defects. First of all, it is not difficult to see that it might not exist. This is in particular the case when a hyperplane separates the two subsets of $\mathbb{R}^d$ given by $\mathcal{W}_+ = \{W_i,\, Y_i = +1\}$ and $\mathcal{W}_- = \{W_i,\, Y_i = -1\}$, i.e. when there exists a unit vector $w_0 \in \mathbb{R}^d$ such that $\langle w, w_0\rangle > 0$ for all $w \in \mathcal{W}_+$ and $\langle w, w_0\rangle < 0$ for $w \in \mathcal{W}_-$. In this case, the conditional likelihood

function at λw_0 with $\lambda > 0$ can be written as

$$\prod_{i=1}^{n} \frac{1}{1+\exp\left[-\lambda Y_i \left\langle w_0, W_i \right\rangle\right]} = \prod_{i=1}^{n} \frac{1}{1+\exp\left[-\lambda \left|\left\langle w_0, W_i \right\rangle\right|\right]} \underset{\lambda \to +\infty}{\longrightarrow} 1,$$

hence the maximal value 1 is not reached. For a thorough study of the existence of the MLE in the logit model we refer to Candès and Sur (2020) as well as the references therein.

Another issue with the use of the MLE lies in the fact that it is not robust and we shall illustrate its instability in our simulation study. Robustness is nevertheless an important property in practice since, going back to Example 2.1.1, it may happen that our database contains a few corrupted data that correspond to mislabelled patients (some patients might have developed a disease which is not D but has similar symptoms) or that the relation (2.1.1) is only approximately true. A natural question arises: how can we provide a suitable estimation of $\mathbb{P}\left[Y = y | W = w\right]$ despite the presence of possibly corrupted data or a slight misspecification of the model?

This is the kind of issue we want to solve here. Our approach is not, however, restricted to the logit model but applies more generally whenever the conditional distribution of Y given W belongs to a one-parameter exponential family as we described in Chapter 1. More precisely, we shall work within the following statistical framework. We observe n pairs of independent, typically non i.i.d., random variables $X_1 = (W_1, Y_1), \ldots, X_n = (W_n, Y_n)$ (with $W_i \in \mathscr{W}$ and $Y_i \in \mathscr{Y}$ for $i \in \{1, \ldots, n\}$) and we want to estimate the n conditional distributions $Q_i^\star(w_i)$ of Y_i when $W_i = w_i$, $i \in \{1, \ldots, n\}$, without any information about the distributions P_{W_i} of the variables W_i which are unknown and can be completely arbitrary. In order to do so, we introduce a statistical model for the $Q_i^\star(w_i)$. We start from an exponential family $\{Q_\theta,\, \theta \in I\}$ where I is an interval of $\mathbb{R}$ and consider the family of conditional distributions $\{Q_{\boldsymbol{\theta}(w)}, \boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}}\}$ where $\overline{\boldsymbol{\Theta}}$ is a given set of functions from $\mathscr{W}$ to I. This provides a model for the n conditional distributions $Q_i^\star(w_i)$ if we pretend that $Q_i^\star(w_i)$ takes the form $Q_{\boldsymbol{\theta}^\star(w_i)}$ for some $\boldsymbol{\theta}^\star \in \overline{\boldsymbol{\Theta}}$, i.e.

$$Q_i^\star(w_i) = Q_{\theta_i^\star} \quad \text{with} \quad \theta_i^\star = \boldsymbol{\theta}^\star(w_i) \quad \text{for } i \in \{1, \ldots, n\}. \tag{2.1.2}$$

We shall do as if (2.1.2) were true although we do not assume it. We merely hope that the set of conditional distributions $Q_{\theta_i^\star}$ induced by a suitable element $\boldsymbol{\theta}^\star$ of $\overline{\boldsymbol{\Theta}}$ provides a reasonably good approximation for the true conditional distributions $Q_i^\star(w_i)$. Any estimator $\widetilde{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}^\star$ leads, by an application of (2.1.2), to an estimator $Q_{\widetilde{\boldsymbol{\theta}}(w_i)}$ of the conditional distribution $Q_i^\star(w_i)$. We measure the risk of such an estimator by a Hellinger-type distance between the conditional distributions $Q_i^\star(w_i)$ and their estimators, integrated with respect to the probabilities P_{W_i} (to be defined in the next section).

Given the model indexed by the elements of $\overline{\boldsymbol{\Theta}}$, instead of estimating $\boldsymbol{\theta}^\star$ by the maximum likelihood method as is commonly done, we use for this a ρ-estimator $\widehat{\boldsymbol{\theta}}$, whose

definition and performance are described in great details in Baraud et al. (2017) and Baraud and Birgé (2018). The purpose of this replacement is to avoid various drawbacks connected to the use of the MLE:
— It may not exist;
— It is typically difficult to evaluate its performance in a non-asymptotic framework and its analysis generally requires some knowledge or restrictions about the distributions of the W_i;
— Its performance may be very bad when the model is not exact (misspecification, presence of outliers, contamination, etc.) as demonstrated by our simulations in Section 2.5.

On the contrary a ρ-estimator always exists and it enjoys the following properties.

- When the parameter set $\overline{\boldsymbol{\Theta}}$ is VC-subgraph with VC-dimension V, the non-asymptotic risk of $\widehat{\boldsymbol{\theta}}$ is bounded by the sum of two terms : an approximation term reflecting the distance between the model and the truth and an estimation term corresponding to the risk bound one would get if the model were true. Moreover, this second term only depends on V. This risk bound involves explicit constants and holds under the only assumption that the data $(W_1, Y_1), \ldots, (W_n, Y_n)$ are independent;

- the estimator $\widehat{\boldsymbol{\theta}}$ still performs well when the function $\boldsymbol{\theta}^\star$ does not belong to $\overline{\boldsymbol{\Theta}}$ but lies close enough to it;

- the estimator is robust: its performance remains stable when the data set $X_1 = (W_1, Y_1), \ldots, X_n = (W_n, Y_n)$ is contaminated or contains outliers or when the statistical model based the exponential family is only approximately correct.

- when the model is exact, the exponential family $\{Q_\theta,\, \theta \in I\}$ is suitably parametrized and $\overline{\boldsymbol{\Theta}}$ is a Hölderian class of smoothness, the estimator $\widehat{\boldsymbol{\theta}}$ is rate optimal (up to a logarithmic factor).

The work presented here is different from the study of ρ-estimators conducted in Baraud and Birgé (2018)[Section 9] for estimating a regression function (seen as the parameter of interest in the conditional distribution of Y given W). In Baraud and Birgé (2018), the authors studied a regression model in which the errors are assumed to be i.i.d., homoscedastic with a density with respect to the Lebesgue measure. In the present paper, the errors are typically heteroscedastic, independent but not i.i.d. and they may not admit a density with respect to the Lebesgue measure. This is the case in the logistic and Poisson regression settings for example. Actually, new results had to be established in order to analyze further the behaviour of ρ-estimators in the statistical setting we consider here. The proof of our main result combines the theory of ρ-estimation — see Baraud et al. (2017) and Baraud and Birgé (2018) — and an original result that establishes the fact

that the family of functions on $\mathscr{W} \times \mathscr{Y}$ of the form

$$(w, y) \mapsto S(y)\boldsymbol{\eta}(w) - A(\boldsymbol{\eta}(w)) \quad \text{with} \quad \boldsymbol{\eta} \in \boldsymbol{\Gamma}$$

is VC-subgraph when S is an arbitrary function on $\mathscr{Y}$, A a convex function defined on an interval I of positive length and $\boldsymbol{\Gamma}$ a VC-subgraph class of functions defined on $\mathscr{W}$ with values in I. The proof of our main result also relies on an upper bound with explicit constants (see Theorem 2.6.1) on the expectation of the supremum of an empirical process over a VC-class of functions. Since we are not aware of such a result (with explicit constants) in the literature, this bound can be of independent interest.

Besides our theoretical guarantees on the performance of the estimator $\widehat{\boldsymbol{\theta}}$, we carry out a simulation study in order to compare it with the MLE and median-based estimators. The simulation study addresses both the situations where the data are generated from the model and when it is contaminated or contains an outlier. To our knowledge, it is the first time that ρ-estimators are implemented numerically and their performance is studied on simulated data.

This remainder of this chapter is organized as follows. We describe our statistical framework in Section 2.2. The construction of the estimator and our main result about its risk are presented in Section 2.3. We also explain why the deviation inequality we derive guarantees the desired robustness property of the estimator. Uniform risk bounds over Hölderian classes are established in Section 2.4 provided that the exponential family involved in the model is suitably parametrized. We also show that, without such a suitable parametrization, the minimax rates may differ from the usual ones established for an homoscedastic Gaussian regression as described by our Example 1.0.1. Section 2.5 is devoted to the description of our algorithm and the simulation study. Our bound on the expectation of the supremum of an empirical process over a VC-subgraph class can be found in Section 2.6 as well as its proof. Section 2.7 is devoted to the other proofs of this chapter.

2.2 The statistical setting

Let us recall that we observe n pairs of independent, but not necessarily i.i.d., random variables $X_1 = (W_1, Y_1), \ldots, X_n = (W_n, Y_n)$ with values in a measurable product space $(\mathscr{X}, \mathcal{X}) = (\mathscr{W} \times \mathscr{Y}, \mathcal{W} \otimes \mathcal{Y})$ and we assume that, for each $i \in \{1, \ldots, n\}$, the conditional distribution of Y_i given $W_i = w_i$ exists and is given by the value at w_i of a measurable function $Q_i^\star$ from $(\mathscr{W}, \mathcal{W})$ to the set $\mathscr{T}$ of all probabilities on $(\mathscr{Y}, \mathcal{Y})$. We equip $\mathscr{T}$ with the Borel σ-algebra $\mathcal{T}$ associated to the total variation distance (which induces the same topology as the Hellinger one defined by (1.3.2)). With this choice of $\mathcal{T}$, the mapping $w \mapsto h^2(Q_i^\star(w), Q)$ on $(\mathscr{W}, \mathcal{W})$ is measurable whatever the probability $Q \in \mathscr{T}$ and $i \in \{1, \ldots, n\}$.

Apart from independence of the W_i, $1 \leq i \leq n$, we assume nothing about their respective distributions P_{W_i} which can therefore be arbitrary.

Let $\overline{\mathscr{Q}} \subset \mathscr{T}$ be an exponential family on the measured space $(\mathscr{Y}, \mathcal{Y}, \nu)$ where ν is an arbitrary σ-finite (positive) measure. We assume that $\overline{\mathscr{Q}} = \{Q_\theta,\ \theta \in I\}$ is indexed by a natural parameter θ that belongs to some interval $I \subset \mathbb{R}$ such that $\mathring{I} \neq \varnothing$. This means that, for all $\theta \in I$, the distribution Q_θ admits a density (with respect to ν) of the form

$$q_\theta : y \mapsto e^{S(y)\theta - A(\theta)} \quad \text{with} \quad A(\theta) = \log \left[\int_{\mathscr{Y}} e^{\theta S(y)} d\nu(y) \right], \tag{2.2.1}$$

where S is a real-valued measurable function on $(\mathscr{Y}, \mathcal{Y})$ which does not coincide with a constant ν-a.e. We also recall that the function A is infinitely differentiable on $\mathring{I}$ and strictly convex on I. It is of course possible to parametrize $\overline{\mathscr{Q}}$ in a different way (i.e. with a non-natural parameter) by performing a variable change $\gamma = v(\theta)$ where v is a continuous and strictly monotone function on I. We shall see in Section 2.3.3 that our main result remains unchanged under such a transformation and we therefore choose, for the sake of simplicity, to introduce it under a natural parametrization first.

Given a class of functions $\overline{\boldsymbol{\Theta}}$ from $\mathscr{W}$ into I, we presume that there exists $\boldsymbol{\theta}^\star$ in $\overline{\boldsymbol{\Theta}}$ such that the conditional distribution $Q_i^\star(w_i)$ is of the form $Q_{\boldsymbol{\theta}^\star(w_i)}$ for all $i \in \{1, \dots, n\}$ and $w_i \in \mathscr{W}$. We refer to $\boldsymbol{\theta}^\star$ as the *regression function*. Even though our estimator is based on these assumptions, we should keep in mind that our statistical model might be misspecified: the conditional distributions $Q_i^\star(w_i)$ might not be exactly of the form $Q_{\boldsymbol{\theta}^\star(w_i)}$, the set $\overline{\boldsymbol{\Theta}}$ might not contain $\boldsymbol{\theta}^\star$ or some observations might be outliers. It will follow from our risk bounds as described by Theorem 2.3.1 that such misspecifications result in an additional term in the risk corresponding to the approximation error between the truth and the model. This term is small when our model provides a good enough approximation of the truth.

For $i \in \{1, \dots, n\}$, let $\mathscr{Q}_\mathscr{W}$ be the set of all measurable mappings (conditional probabilities) from $(\mathscr{W}, \mathcal{W})$ into $(\mathscr{T}, \mathcal{T})$. We set $\boldsymbol{\mathcal{Q}}_\mathscr{W} = \mathscr{Q}_\mathscr{W}^n$ so that the n-tuple $\mathbf{Q}^\star = (Q_1^\star, \dots, Q_n^\star)$ belongs to $\boldsymbol{\mathcal{Q}}_\mathscr{W}$ as well as the n-tuple $\mathbf{Q}_{\boldsymbol{\theta}} = (Q_{\boldsymbol{\theta}}, \dots, Q_{\boldsymbol{\theta}})$ where $Q_{\boldsymbol{\theta}} \in \mathscr{Q}_\mathscr{W}$ denotes the mapping $w \mapsto Q_{\boldsymbol{\theta}(w)}$ when $\boldsymbol{\theta}$ is a measurable function from $\mathscr{W}$ into I. We endow the space $\boldsymbol{\mathcal{Q}}_\mathscr{W}$ with the Hellinger-type (pseudo) distance $\mathbf{h}$ defined as follows. For $\mathbf{Q} = (Q_1, \dots, Q_n)$ and $\mathbf{Q}' = (Q_1', \dots, Q_n')$ in $\boldsymbol{\mathcal{Q}}_\mathscr{W}$,

$$\begin{aligned} \mathbf{h}^2(\mathbf{Q}, \mathbf{Q}') &= \mathbb{E}\left[\sum_{i=1}^n h^2\left(Q_i(W_i), Q_i'(W_i)\right) \right] \\ &= \sum_{i=1}^n \int_\mathscr{W} h^2\left(Q_i(w), Q_i'(w)\right) dP_{W_i}(w). \end{aligned} \tag{2.2.2}$$

In particular, $\mathbf{h}(\mathbf{Q}, \mathbf{Q}') = 0$ implies that for all $i \in \{1, \dots, n\}$, $Q_i = Q_i'$ P_{W_i}-a.s.

On the basis of the observations $X_1, \ldots, X_n$, we build an estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}^\star$ with values in $\overline{\boldsymbol{\Theta}}$ and evaluate its performance by the quantity

$$\mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_{\widehat{\boldsymbol{\theta}}}) = \sum_{i=1}^{n} \int_{\mathscr{W}} h^2\left(Q_i^\star(w), Q_{\widehat{\boldsymbol{\theta}}(w)}\right) dP_{W_i}(w).$$

When P is the distribution of a random variable $(W, Y) \in \mathscr{W} \times \mathscr{Y}$ we write it as $P = Q \cdot P_W$ where P_W is the marginal distribution of W and Q the conditional distribution of Y given W. For $P = Q \cdot P_W$ and $P' = Q' \cdot P_W$ the squared Hellinger distance between P and P' is written as

$$h^2(P, P') = \int_{\mathscr{W}} h^2\left(Q(w), Q'(w)\right) dP_W(w).$$

Setting, for $i \in \{1, \ldots, n\}$ and $\boldsymbol{\theta}$ a function from $\mathscr{W}$ to I, $P_i^\star = Q_i^\star \cdot P_{W_i}$ and $P_{i,\boldsymbol{\theta}} = Q_{\boldsymbol{\theta}} \cdot P_{W_i}$ we deduce that

$$\mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_{\boldsymbol{\theta}}) = \sum_{i=1}^{n} h^2\left(P_i^\star, P_{i,\boldsymbol{\theta}}\right)$$

so that $\mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_{\boldsymbol{\theta}})$ is equal to $\mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\boldsymbol{\theta}}) = \sum_{i=1}^{n} h^2\left(P_i^\star, P_{i,\boldsymbol{\theta}}\right)$ where $\mathbf{P}^\star = \otimes_{i=1}^{n} P_i^\star$ is the true distribution of the observed data $\boldsymbol{X} = (X_1, \ldots, X_n)$ while $\mathbf{P}_{\boldsymbol{\theta}} = \bigotimes_{i=1}^{n} P_{i,\boldsymbol{\theta}} = \bigotimes_{i=1}^{n} (Q_{\boldsymbol{\theta}} \cdot P_{W_i})$ is the joint distribution of independent random variables (W_i', Y_i') with $1 \leq i \leq n$ for which the conditional distribution of Y_i' given $W_i' = w_i$ is given by $Q_{\boldsymbol{\theta}(w_i)} \in \overline{\mathscr{Q}}$ for all i. This shows that the quantity $\mathbf{h}(\mathbf{Q}^\star, \mathbf{Q}_{\boldsymbol{\theta}}) = \mathbf{h}(\mathbf{P}^\star, \mathbf{P}_{\boldsymbol{\theta}})$ may also be interpreted as a distance between the probability distributions $\mathbf{P}^\star$ and $\mathbf{P}_{\boldsymbol{\theta}}$ and not only as a (pseudo) distance between the conditional ones $\mathbf{Q}^\star$ and $\mathbf{Q}_{\boldsymbol{\theta}}$. More generally, given two measurable functions $\boldsymbol{\theta}, \boldsymbol{\theta}'$ from $\mathscr{W}$ to I, the quantity $\mathbf{h}(\mathbf{Q}_{\boldsymbol{\theta}}, \mathbf{Q}_{\boldsymbol{\theta}'})$ can also be written as $\mathbf{h}(\mathbf{P}_{\boldsymbol{\theta}}, \mathbf{P}_{\boldsymbol{\theta}'})$. Note that, unlike $\mathbf{Q}_{\widehat{\boldsymbol{\theta}}}$, $\mathbf{P}_{\widehat{\boldsymbol{\theta}}}$ is not an estimator (of $\mathbf{P}^\star$) since it depends on the marginal distributions $P_{W_1}, \ldots, P_{W_n}$ which are unknown.

2.2.1 Examples

Let us present here some typical statistical models to which our approach applies.

Example 2.2.1 (Homoscedastic Gaussian regression with known variance)**.** Given n independent random variables $W_1, \ldots, W_n$ with values in $\mathscr{W}$, let

$$Y_i = \boldsymbol{\theta}^\star(W_i) + \sigma\varepsilon_i \quad \text{for all } i \in \{1, \ldots, n\},$$

where the ε_i are i.i.d. standard real-valued Gaussian random variables, σ is a known positive number and $\boldsymbol{\theta}^\star$ an unknown regression function with values in $I = \mathbb{R}$. In this case, $\overline{\mathscr{Q}}$ is the set of all Gaussian distributions with variance σ^2 and for all $\theta \in I = \mathbb{R}$, $Q_\theta = \mathcal{N}(\theta, \sigma^2)$ has a density with respect to $\nu = \mathcal{N}(0, \sigma^2)$ on $(\mathscr{Y}, \mathcal{Y}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ which is of the form (2.2.1) with $A(\theta) = \theta^2/(2\sigma^2)$ and $S(y) = y/\sigma^2$ for all $y \in \mathbb{R}$.

Example 2.2.2 (Binary regression)**.** The pairs of random variables (W_i, Y_i) with $i \in \{1, \dots, n\}$ are independent with values in $\mathscr{W} \times \{0, 1\}$ and

$$\mathbb{P}[Y_i = y | W_i = w_i] = \frac{\exp[y\boldsymbol{\theta}^\star(w_i)]}{1 + \exp[\boldsymbol{\theta}^\star(w_i)]} \quad \text{for all } y \in \{0, 1\} \text{ and } w_i \in \mathscr{W}. \tag{2.2.3}$$

This means that the conditional distribution of Y_i given $W_i = w_i$ is Bernoulli with mean $(1 + \exp[-\boldsymbol{\theta}^\star(w_i)])^{-1}$ for some regression function $\boldsymbol{\theta}^\star$ with values in $I = \mathbb{R}$. This model is equivalent to the logit one presented in Example 2.1.1 by changing $Y_i \in \{0, 1\}$ into $Y_i' = 2Y_i - 1 \in \{-1, 1\}$ for all i. The exponential family $\overline{\mathscr{Q}}$ consists of the Bernoulli distribution Q_θ with mean $1/[1 + e^{-\theta}] \in (0, 1)$ and $\theta \in I = \mathbb{R}$. For all $\theta \in \mathbb{R}$, Q_θ admits a density with respect to the counting measure ν on $\mathscr{Y} = \{0, 1\}$ of the form (2.2.1) with $A(\theta) = \log(1 + e^\theta)$ and $S(y) = y$ for all $y \in \mathscr{Y}$.

Example 2.2.3 (Poisson regression)**.** The exponential family $\overline{\mathscr{Q}}$ is the set of all Poisson distributions Q_θ with mean e^θ, $\theta \in I = \mathbb{R}$. Taking for ν the Poisson distribution with mean 1, the density of Q_θ with respect to ν takes the form (2.2.1) with $S(y) = y$ for all $y \in \mathbb{N}$ and $A(\theta) = e^\theta - 1$ for all $\theta \in \mathbb{R}$. The conditional distribution of Y_i given $W_i = w_i$ is presumed to be Poisson with mean $\exp[\boldsymbol{\theta}^\star(w_i)]$ for some regression function $\boldsymbol{\theta}^\star$ with values in $I = \mathbb{R}$.

Example 2.2.4 (Exponential multiplicative regression)**.** The random variables W_i are independent and

$$Y_i = \frac{Z_i}{\boldsymbol{\theta}^\star(W_i)} \quad \text{for all } i \in \{1, \dots, n\} \tag{2.2.4}$$

where the Z_i are i.i.d. with exponential distribution of parameter 1 and independent of the W_i. The conditional distribution of Y_i given $W_i = w_i$ is then exponential with mean $1/\boldsymbol{\theta}^\star(w_i) \in I = (0, +\infty)$. Exponential distributions parametrized by $\theta \in I$ admit densities with respect to the Lebesgue measure on $\mathbb{R}_+$ of the form (2.2.1) with $S(y) = -y$ for all $y \in \mathscr{Y} = \mathbb{R}_+$ and $A(\theta) = -\log\theta$.

2.3 The main results

2.3.1 The estimation procedure

As mentioned in the introduction, our approach is based on ρ-estimation. The basic ideas that underline the construction of these estimators have been explained in Section 1.4.2 and more details can be found in Baraud and Birgé (2018). Let ψ be the function defined on $[0, +\infty]$ by

$$\psi(x) = \frac{x - 1}{x + 1} \quad \text{for } x \in [0, +\infty) \quad \text{and} \quad \psi(+\infty) = 1. \tag{2.3.1}$$

Let us set, for $\boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}}$, $q_{\boldsymbol{\theta}}(X_i) = q_{\boldsymbol{\theta}(W_i)}(Y_i)$, where q_θ is given by (2.2.1) and, in order to avoid measurability issues, let us restrict ourselves to those $\boldsymbol{\theta}$ belonging to a finite or countable subset $\boldsymbol{\Theta}$ of $\overline{\boldsymbol{\Theta}}$. We then introduce the function

$$\mathbf{T}(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\theta}') = \sum_{i=1}^{n} \psi\left(\sqrt{\frac{q_{\boldsymbol{\theta}'}(X_i)}{q_{\boldsymbol{\theta}}(X_i)}}\right) \quad \text{for} \quad \boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}, \tag{2.3.2}$$

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

$$\boldsymbol{\upsilon}(\boldsymbol{X}, \boldsymbol{\theta}) = \sup_{\boldsymbol{\theta}' \in \boldsymbol{\Theta}} \mathbf{T}(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\theta}') \quad \text{for all} \quad \boldsymbol{\theta} \in \boldsymbol{\Theta} \tag{2.3.3}$$

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

$$\mathscr{E}(\boldsymbol{X}) = \left\{\boldsymbol{\theta} \in \boldsymbol{\Theta} \text{ such that } \boldsymbol{\upsilon}(\boldsymbol{X}, \boldsymbol{\theta}) \leq \inf_{\boldsymbol{\theta}' \in \boldsymbol{\Theta}} \boldsymbol{\upsilon}(\boldsymbol{X}, \boldsymbol{\theta}') + \frac{\kappa_\rho}{25}\right\} \tag{2.3.4}$$

with $\kappa_\rho = 280\sqrt{2} + 74$, so that $18 < \kappa_\rho/25 < 18.8$. The random variable $\widehat{\boldsymbol{\theta}}(\boldsymbol{X})$ is our estimator of the regression function $\boldsymbol{\theta}^\star$ and $\mathbf{Q}_{\widehat{\boldsymbol{\theta}}} = (Q_{\widehat{\boldsymbol{\theta}}}, \ldots, Q_{\widehat{\boldsymbol{\theta}}})$.

Note that the construction of the estimator is only based on the choices of the exponential family given by (2.2.1) and the subset $\boldsymbol{\Theta}$ of $\overline{\boldsymbol{\Theta}}$. In particular, the estimator does not depend on the distributions P_{W_i} of the W_i which may therefore be unknown.

The fact that we build our estimator on a finite or countable subset $\boldsymbol{\Theta}$ of $\overline{\boldsymbol{\Theta}}$ is not restrictive as we shall see. Besides, this assumption is consistent with the practice of calculating an estimator on a computer that can handle a finite number of values only.

Let us mention that similar results could be established for the ρ-estimator associated to the alternative choice $\psi(x) = (x-1)/\sqrt{x^2+1}$. Nevertheless, the risk bounds we get for this choice of ψ involve numerical constants that are larger than those we establish here for $\psi(x) = (x-1)/(x+1)$. We therefore focus on this latter choice of ψ.

2.3.2 The main assumption and the performance of $\widehat{\boldsymbol{\theta}}$

Our main assumption is stated as follows.

Assumption 2.3.1. The class of functions $\overline{\boldsymbol{\Theta}}$ is VC-subgraph on $\mathscr{W}$ with dimension not larger than $V \geq 1$.

We refer to Section 1.5 for the definition of VC-subgraph classes and their properties. In this chapter, we mainly use the facts that Assumption 2.3.1 is satisfied when $\overline{\boldsymbol{\Theta}}$ is a linear space $\mathcal{V}$ with finite dimension $d \geq 1$, in which case $V = d+1$ by Proposition 1.5.1 and that it is also satisfied when $\overline{\boldsymbol{\Theta}}$ is of the form $\{F(\boldsymbol{\beta}),\ \boldsymbol{\beta} \in \mathcal{V}\}$ where F is a monotone function on the real-line. In this latter case, the VC-dimension of $\overline{\boldsymbol{\Theta}}$ is not larger than that of $\mathcal{V}$ according to Proposition 1.5.2. We set

$$c_1 = 150, \quad c_2 = 1.1 \times 10^6, \quad c_3 = 5014 \tag{2.3.5}$$

and, for $\mathbf{Q} \in \mathscr{Q}_{\mathscr{W}}$ and $\mathbf{A} \subset \mathscr{Q}_{\mathscr{W}}$,

$$\mathbf{h}\left(\mathbf{Q}, \mathbf{A}\right) = \inf_{\mathbf{Q}' \in \mathbf{A}} \mathbf{h}\left(\mathbf{Q}, \mathbf{Q}'\right).$$

The following theorem provides a probabilistic bound for the distance between our estimator $\mathbf{Q}_{\widehat{\boldsymbol{\theta}}}$ and $\mathbf{Q}^\star$.

Theorem 2.3.1. *Let $\xi > 0$. Under Assumption 2.3.1, whatever the conditional probabilities $\mathbf{Q}^\star = (Q_1^\star, \ldots, Q_n^\star)$ of the Y_i given W_i and the distributions of the W_i, the estimator $\mathbf{Q}_{\widehat{\boldsymbol{\theta}}}$ defined in Section 2.3.1 satisfies, with a probability at least $1 - e^{-\xi}$,*

$$\mathbf{h}^2\left(\mathbf{Q}^\star, \mathbf{Q}_{\widehat{\boldsymbol{\theta}}}\right) \leq c_1 \mathbf{h}^2(\mathbf{Q}^\star, \mathscr{Q}) + c_2 V\left[9.11 + \log_+\left(\frac{n}{V}\right)\right] + c_3\left(1.5 + \xi\right) \tag{2.3.6}$$

where $\mathscr{Q} = \{\mathbf{Q}_{\boldsymbol{\theta}} = (Q_{\boldsymbol{\theta}}, \ldots, Q_{\boldsymbol{\theta}}),\ \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ and $\log_+ = \max(0, \log)$.

The constants c_1, c_2 and c_3 are numerical constants. They are independent of the choice of the exponential family. When the model $\mathscr{Q}$ is exact, the bound we get only depends on the VC-dimension of $\overline{\boldsymbol{\Theta}}$.

It is clear that (2.3.6) also holds true for $\overline{\mathscr{Q}} = \{\mathbf{Q}_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}}\}$ in place of $\mathscr{Q}$ when $\mathscr{Q}$ is dense in $\overline{\mathscr{Q}}$ with respect to the Hellinger-type distance $\mathbf{h}$. This is the case when $\boldsymbol{\Theta}$ is dense in $\overline{\boldsymbol{\Theta}}$ for the topology of pointwise convergence. We do not comment on our result any further in this direction and rather refer to Baraud and Birgé (2018) Section 4.2. From now on, we assume for the sake of simplicity that $\mathscr{Q}$ is dense in $\overline{\mathscr{Q}}$, doing as if $\overline{\boldsymbol{\Theta}} = \boldsymbol{\Theta}$. In the remaining part of this section, C will denote a positive numerical constant that may vary from line to line.

Let us now rewrite (2.3.6) in a slightly different form. We have seen in Section 2.2 that the quantity $\mathbf{h}\left(\mathbf{Q}^\star, \mathbf{Q}_{\boldsymbol{\theta}}\right)$ with $\boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}}$, which involves the conditional probabilities of $\mathbf{P}^\star$ and $\mathbf{P}_{\boldsymbol{\theta}}$ with respect to the W_i, can also be interpreted in terms of the Hellinger(-type) distance between these two product probabilities. Inequality (2.3.6) therefore implies that

$$\mathbb{P}\left[C\mathbf{h}^2\left(\mathbf{P}^\star, \mathbf{P}_{\widehat{\boldsymbol{\theta}}}\right) > \mathbf{h}^2(\mathbf{P}^\star, \overline{\mathscr{P}}) + V\left[1 + \log_+\left(\frac{n}{V}\right)\right] + \xi\right] \leq e^{-\xi}, \tag{2.3.7}$$

where $\overline{\mathscr{P}} = \{\mathbf{P}_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}}\}$. Integrating this inequality with respect to $\xi > 0$ leads to the following risk bound for our estimator $\widehat{\boldsymbol{\theta}}$

$$C\mathbb{E}\left[\mathbf{h}^2\left(\mathbf{P}^\star, \mathbf{P}_{\widehat{\boldsymbol{\theta}}}\right)\right] \leq \mathbf{h}^2(\mathbf{P}^\star, \overline{\mathscr{P}}) + V\left[1 + \log_+\left(\frac{n}{V}\right)\right]. \tag{2.3.8}$$

In order to comment upon (2.3.8), let us start with the ideal situation where $\mathbf{P}^\star$ belongs to $\overline{\mathscr{P}}$, i.e. $\mathbf{P}^\star = \mathbf{P}_{\boldsymbol{\theta}^\star}$ for some $\boldsymbol{\theta}^\star \in \overline{\boldsymbol{\Theta}}$, in which case (2.3.8) leads to

$$C\mathbb{E}\left[\mathbf{h}^2\left(\mathbf{P}_{\boldsymbol{\theta}^\star}, \mathbf{P}_{\widehat{\boldsymbol{\theta}}}\right)\right] \leq V\left[1 + \log_+\left(\frac{n}{V}\right)\right]. \tag{2.3.9}$$

Up to the logarithmic factor, the right-hand side of this inequality is of the expected order of magnitude V for the quantity $\mathbf{h}^2(\mathbf{P}_{\boldsymbol{\theta}^\star}, \mathbf{P}_{\widehat{\boldsymbol{\theta}}})$: in typical situations V is of the same order as the number of parameters that are used to parametrize $\overline{\boldsymbol{\Theta}}$.

When the true distribution $\mathbf{P}^\star$ is of the form $\mathbf{P}_{\boldsymbol{\theta}^\star}$ but the regression function $\boldsymbol{\theta}^\star$ does not belong to $\overline{\boldsymbol{\Theta}}$, or when the conditional distributions of the Y_i given W_i do not belong to our exponential family, inequality (2.3.8) shows that, as compared to (2.3.9), the bound we get involves the approximation term $\mathbf{h}^2(\mathbf{P}^\star, \overline{\mathscr{P}})$ that accounts for the fact that our statistical model is misspecified. However, as long as this quantity remains small enough as compared to $V\left[1+\log_+(n/V)\right]$, our risk bound will be of the same order as that given by (2.3.9) when the model is exact. This property accounts for the stability of our estimation procedure under misspecification. In order to be more specific, let us assume that

$$\mathbf{P}^\star = \bigotimes_{i=1}^n \left[(1-\alpha_i)P_{i,\overline{\boldsymbol{\theta}}} + \alpha_i R_i\right] \quad \text{and} \quad \sum_{i=1}^n \alpha_i \leq \frac{n}{2} \tag{2.3.10}$$

where $\overline{\boldsymbol{\theta}} \in \overline{\boldsymbol{\Theta}}$, R_i is an arbitrary distribution on $\mathscr{X}$ and α_i a number in $[0,1]$ for all $i \in \{1,\dots,n\}$. Such a distribution $\mathbf{P}^\star$ allows us to model different form of robustness including robustness to the presence of contaminating data as well as outliers. In the case of contamination, $P_{W_i} = P_W$, $\alpha_i = \alpha \in (0, 1/2]$, $R_i = R \neq P_{\overline{\boldsymbol{\theta}}} = Q_{\overline{\boldsymbol{\theta}}} \cdot P_W$ for all $i \in \{1,\dots,n\}$ and one observes an n-sample a portion $(1-\alpha)$ of which is drawn according to a distribution $P_{\overline{\boldsymbol{\theta}}}$ that belongs to our model $\overline{\mathscr{P}} = \{P_{\boldsymbol{\theta}} = Q_{\boldsymbol{\theta}} \cdot P_W,\ \boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}}\}$ while the remaining part of the data is drawn according to a contaminating distribution R. In the second case, the data set contains the outliers $\{a_i,\ i \in K\}$ for some subset $K \subset \{1,\dots,n\}$ with $K \neq \varnothing$ so that $\mathbf{P}^\star$ is of the form (2.3.10) with $\alpha_i = \mathbb{1}_{i\in K}$ for all $i \in \{1,\dots,n\}$ and $R_i = \delta_{a_i}$ for all $i \in K$. In all cases, using the classical inequality $h^2 \leq D$ where D denotes the total variation distance between probabilities, we get

$$\mathbf{h}^2(\mathbf{P}^\star, \overline{\mathscr{P}}) \leq \mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\overline{\boldsymbol{\theta}}}) \leq \sum_{i=1}^n D(P_i^\star, P_{\overline{\boldsymbol{\theta}}}) \leq \sum_{i=1}^n \alpha_i, \tag{2.3.11}$$

which means that whenever $\sum_{i=1}^n \alpha_i$ remains small as compared to $V(1+\log_+(n/V))$, the performance of the estimator remains almost the same as if $\mathbf{P}^\star$ were equal to $\mathbf{P}_{\overline{\boldsymbol{\theta}}}$. The estimator $\widehat{\boldsymbol{\theta}}$ therefore possesses some stability properties with respect to contamination and the presence of outliers.

2.3.3 From a natural to a general exponential family

So far we focused on an exponential family $\overline{\mathscr{Q}}$ parametrized by its natural parameter. However statisticians often write exponential families $\overline{\mathscr{Q}}$ under the general form $\overline{\mathscr{Q}} = \{R_\gamma = r_\gamma \cdot \nu,\ \gamma \in J\}$ with

$$r_\gamma : y \mapsto e^{u(\gamma)S(y) - B(\gamma)} \quad \text{for } \gamma \in J. \tag{2.3.12}$$

In (2.3.12), J denotes a (non-degenerate) interval of $\mathbb{R}$ and u a continuous and strictly monotone function from J onto I so that $B = A \circ u$. In the exponential family $\overline{\mathscr{Q}} =$

$\{R_\gamma,\ \gamma \in J\} = \{Q_\theta,\ \theta \in I\}$, the probabilities R_γ are associated to the probabilities Q_θ by the formula $R_\gamma = Q_{u(\gamma)}$.

With this new parametrization, we could alternatively write our statistical model $\overline{\mathscr{Q}}$ as

$$\overline{\mathscr{Q}} = \left\{\mathbf{R}_{\boldsymbol{\gamma}} = (R_{\boldsymbol{\gamma}}, \ldots, R_{\boldsymbol{\gamma}}),\ \boldsymbol{\gamma} \in \overline{\boldsymbol{\Gamma}}\right\} \tag{2.3.13}$$

where $\overline{\boldsymbol{\Gamma}}$ is a class of functions $\boldsymbol{\gamma}$ from $\mathscr{W}$ into J. Starting from such a statistical model and presuming that $\mathbf{Q}^\star = \mathbf{R}_{\boldsymbol{\gamma}^\star}$ for some function $\boldsymbol{\gamma}^\star \in \overline{\boldsymbol{\Gamma}}$, we can build an estimator $\widehat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}^\star$ as follows: given a finite or countable subset $\boldsymbol{\Gamma}$ of $\overline{\boldsymbol{\Gamma}}$ we set $\widehat{\boldsymbol{\gamma}} = u^{-1}(\widehat{\boldsymbol{\theta}})$ where $\widehat{\boldsymbol{\theta}}$ is any estimator obtained by applying the procedure described in Section 2.3.1 under the natural parametrization of the exponential family $\overline{\mathscr{Q}}$ and using the finite or countable model $\boldsymbol{\Theta} = \{\boldsymbol{\theta} = u \circ \boldsymbol{\gamma},\ \boldsymbol{\gamma} \in \boldsymbol{\Gamma}\}$.

Since our model $\overline{\mathscr{Q}}$ for the conditional probabilities $\mathbf{Q}^\star$ is unchanged (only its parametrisation changes), it would be interesting to establish a result on the performance of the estimator $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}} = \mathbf{Q}_{\widehat{\boldsymbol{\theta}}}$ which is independent of the parametrization. A nice feature of the VC-subgraph property lies in the fact that by Proposition 1.5.2, it is preserved by composition with a monotone function: since u is monotone, if $\overline{\boldsymbol{\Gamma}}$ is VC-subgraph with dimension not larger than V, so is $\overline{\boldsymbol{\Theta}}$ and our Theorem 2.3.1 applies. The following corollary is therefore straightforward.

Corollary 2.3.1. *Let $\xi > 0$. If the statistical model $\overline{\mathscr{Q}}$ is under the general form (2.3.13) and $\overline{\boldsymbol{\Gamma}}$ is VC-subgraph with dimension not larger than $V \geq 1$, whatever the conditional probabilities $\mathbf{Q}^\star = (Q_1^\star, \ldots, Q_n^\star)$ of the Y_i given W_i and the distributions of the W_i, the estimator $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}}$ satisfies with a probability at least $1 - e^{-\xi}$,*

$$\mathbf{h}^2\left(\mathbf{Q}^\star, \mathbf{R}_{\widehat{\boldsymbol{\gamma}}}\right) \leq c_1 \mathbf{h}^2(\mathbf{Q}^\star, \mathscr{Q}) + c_2 V\left[9.11 + \log_+\left(\frac{n}{V}\right)\right] + c_3\left(1.5 + \xi\right) \tag{2.3.14}$$

where $\mathscr{Q} = \{\mathbf{R}_{\boldsymbol{\gamma}},\ \boldsymbol{\gamma} \in \boldsymbol{\Gamma}\}$. In particular,

$$\mathbb{E}\left[\mathbf{h}^2\left(\mathbf{Q}^\star, \mathbf{R}_{\widehat{\boldsymbol{\gamma}}}\right)\right] \leq C'\left[\mathbf{h}^2(\mathbf{Q}^\star, \mathscr{Q}) + V\left[1 + \log_+\left(\frac{n}{V}\right)\right]\right], \tag{2.3.15}$$

for some numerical constant $C' > 0$.

A nice feature of our approach lies in the fact that (2.3.14) holds for all exponential families simultaneously and all ways of parametrizing them. In particular, the VC-dimension associated to the model $\overline{\mathscr{Q}}$ is intrinsic since it is independent of the way it is parametrized.

2.4 Uniform risk bounds

Throughout this section, we assume that the W_i are i.i.d. with common distribution P_W and that $\mathbf{Q}^\star = \mathbf{R}_{\boldsymbol{\gamma}^\star} = (R_{\boldsymbol{\gamma}^\star}, \ldots, R_{\boldsymbol{\gamma}^\star})$ belongs to a statistical model of the (general) form given by (2.3.13) where $\overline{\boldsymbol{\Gamma}}$ is a class of smooth functions. More precisely, we assume that

for some $\alpha \in (0,1]$ and $M > 0$, $\overline{\mathbf{\Gamma}} = \mathcal{H}_\alpha(M)$ is the set of functions $\boldsymbol{\gamma}$ on $\mathscr{W} = [0,1]$ with values in J that satisfy the Hölder condition

$$|\boldsymbol{\gamma}(x) - \boldsymbol{\gamma}(y)| \leq M|x-y|^\alpha \quad \text{for all } x, y \in [0,1] \tag{2.4.1}$$

and denote by $\overline{\mathscr{Q}}[\mathcal{H}_\alpha(M)]$ the corresponding family of conditional distributions $Q^\star = R_{\boldsymbol{\gamma}^\star}$. Because of our equidistribution assumption and the form of our loss function, the loss $\mathbf{h}^2\left(\mathbf{Q}^\star, \mathbf{R}_{\widetilde{\boldsymbol{\gamma}}}\right)$ takes the form $nh^2\left(Q^\star, R_{\widetilde{\boldsymbol{\gamma}}}\right)$ whatever the estimator $\widetilde{\boldsymbol{\gamma}}$.

Our aim is both to estimate $\mathbf{R}_{\boldsymbol{\gamma}^\star}$, or equivalently $R_{\boldsymbol{\gamma}^\star}$, under the assumption that $\boldsymbol{\gamma}^\star \in \mathcal{H}_\alpha(M)$ and to evaluate the minimax risk over $\overline{\mathscr{Q}}[\mathcal{H}_\alpha(M)]$, i.e. the quantity

$$\mathcal{R}_n(\mathcal{H}_\alpha(M)) = \inf_{\widetilde{\boldsymbol{\gamma}}} \sup_{\boldsymbol{\gamma}^\star \in \mathcal{H}_\alpha(M)} \mathbb{E}\left[h^2\left(R_{\boldsymbol{\gamma}^\star}, R_{\widetilde{\boldsymbol{\gamma}}}\right)\right] \tag{2.4.2}$$

with

$$h^2\left(R_{\boldsymbol{\gamma}}, R_{\boldsymbol{\gamma}'}\right) \stackrel{\text{def}}{=} \int_{\mathscr{W}} h^2\left(R_{\boldsymbol{\gamma}(w)}, R_{\boldsymbol{\gamma}'(w)}\right) dP_W(w).$$

where the infimum runs among all estimators $\widetilde{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}^\star$ based on the n-sample $X_1, \ldots, X_n$.

2.4.1 Uniform risk bounds over Hölder classes

It is common to parametrize the exponential family $\overline{\mathscr{Q}} = \{R_\gamma,\ \gamma \in J\}$ by the means of the distributions, i.e. with $\gamma = \int_{\mathscr{Y}} y dR_\gamma(y)$. This is typically the case for the Bernoulli, Gaussian and Poisson families for example. In such a case, one can write the model in a heteroscedastic regression form:

$$Y_i = \boldsymbol{\gamma}^\star(W_i) + \sigma\left(\boldsymbol{\gamma}^\star(W_i)\right)\varepsilon_i \quad \text{for all } i \in \{1, \ldots, n\}, \tag{2.4.3}$$

where $\sigma^2\left(\boldsymbol{\gamma}^\star(W_i)\right)$ is the conditional variance of Y_i and the ε_i errors which, conditionally to the W_i are centred and with variance 1. As mentioned in Chapter 1, many authors have used this form and its similarity to the classical Gaussian regression framework given in Example 2.2.1 to derive estimators with performances that mimic those of the least squares estimators in the Gaussian case. In particular, when $\boldsymbol{\gamma}^\star \in \mathcal{H}_\alpha(M)$, the minimax rate for Gaussian regression is $n^{-2\alpha/(2\alpha+1)}$ and, for the more general situation described in (2.4.3) various authors established similar rates for their estimators by using the $\mathbb{L}_2$-loss (under somewhat restrictive assumptions as mentioned in Chapter 1).

With our Hellinger-type loss, we also show in this section that $n^{-2\alpha/(2\alpha+1)}$ is the minimax rate for estimating $R_{\boldsymbol{\gamma}^\star}$ with $\boldsymbol{\gamma}^\star \in \mathcal{H}_\alpha(M)$. However, this result holds when the parametrization of the exponential family satisfies some suitable conditions. When these conditions are not met, the minimax rate may be different as we shall see, even when the exponential family is parametrized by the mean as it is commonly done in the literature. In any case, our estimator is proven to achieve the minimax rate up to a logarithmic factor.

Let us first introduce the following assumptions on the parametrization.

Assumption 2.4.1. There exists a constant $\kappa > 0$ such that

$$h(R_\gamma, R_{\gamma'}) \leq \kappa \left|\gamma - \gamma'\right| \quad \text{for all } \gamma, \gamma' \in J \tag{2.4.4}$$

and for a (non-degenerate) compact interval $K \subset J$, there exists a constant $c_K > 0$ such that

$$h(R_\gamma, R_{\gamma'}) \geq c_K \left|\gamma - \gamma'\right| \quad \text{for all } \gamma, \gamma' \in K. \tag{2.4.5}$$

This assumption is in particular satisfied in the following situation.

Proposition 2.4.1. *Let $\overline{\mathscr{Q}} = \{Q_\theta,\ \theta \in I\}$ be a natural exponential family defined by (2.2.1) where I is an open interval. If the function v satisfies*

$$v'(\theta) = \sqrt{\frac{A''(\theta)}{8}} > 0 \quad \textit{for all } \theta \in I, \tag{2.4.6}$$

when parametrized by $\gamma = v(\theta)$, the exponential family $\overline{\mathscr{Q}} = \{R_\gamma = Q_{v^{-1}(\gamma)},\ \gamma \in J\}$ satisfies Assumption 2.4.1 with $\kappa = 1$ for all choices of a (non-trivial) compact subset K of J.

It is well-known that the functions $v_j(\theta)$, $j \in \{1, 2, 3, 4\}$ given by

$$v_1(\theta) = \frac{\theta}{\sigma\sqrt{8}},\quad v_2(\theta) = \frac{1}{\sqrt{2}} \arcsin\left(\frac{1}{\sqrt{1+e^{-\theta}}}\right),\quad v_3(\theta) = \frac{1}{\sqrt{2}} e^{\theta/2} \quad \text{on } \mathbb{R}$$

and $v_4(\theta) = 8^{-1/2} \log\theta$ on $(0, +\infty)$ satisfy (2.4.6) in the cases of Examples 2.2.1, 2.2.2, 2.2.3 and 2.2.4 respectively.

As a consequence of Assumption 2.4.1 we derive by integration with respect to P_W that, for all functions $\boldsymbol{\gamma}, \boldsymbol{\gamma}'$ on $\mathscr{W}$ with values in J,

$$h^2\left(R_{\boldsymbol{\gamma}}, R_{\boldsymbol{\gamma}'}\right) \leq \kappa^2 \left\|\boldsymbol{\gamma} - \boldsymbol{\gamma}'\right\|_2^2 = \kappa^2 \int_{\mathscr{W}} \left(\boldsymbol{\gamma} - \boldsymbol{\gamma}'\right)^2 dP_W, \tag{2.4.7}$$

which leads to the following uniform risk bound for the performance of the ρ-estimator $R_{\widehat{\boldsymbol{\gamma}}}$ when $\boldsymbol{\gamma}^\star \in \mathcal{H}_\alpha(M)$. Note that this upper bound holds without any assumption on P_W.

Proposition 2.4.2. *Assume that (2.4.4) is satisfied. Let $\alpha \in (0, 1]$, $M > 0$ and $\overline{\boldsymbol{S}}$ be the set of functions with values in the interval J which are piecewise constant on each element of a partition $\{I_j, j \in \{1, \ldots, D\}\}$ of $[0, 1]$ into $D \geq 1$ intervals of lengths $1/D$. For*

$$D = D(\alpha, M, n) = \min\left\{k \in \mathbb{N},\ \left(\frac{\kappa^2 M^2 n}{1 + \log n}\right)^{\frac{1}{1+2\alpha}} \leq k\right\},$$

the ρ-estimator $\widehat{\boldsymbol{\gamma}}$ based on (any) countable and dense subset $\boldsymbol{S}$ of $\overline{\boldsymbol{S}}$ (with respect to the supremum norm) satisfies

$$\sup_{\boldsymbol{\gamma}^\star \in \mathcal{H}_\alpha(M)} \mathbb{E}\left[h^2\left(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}}\right)\right] \leq 2C' \left[\left(\frac{(\kappa M)^{1/\alpha} \log(en)}{n}\right)^{\frac{2\alpha}{1+2\alpha}} + \frac{3\log(en)}{2n}\right]$$

where C' is the numerical constant appearing in (2.3.15).

To show that this rate is optimal under Assumption 2.4.1 and the ρ-estimator minimax (up to a logarithmic factor) when the distribution of the W_i can be arbitrary, let us assume that P_W is the uniform distribution on $\mathscr{W} = [0,1]$. It then follows from (2.4.5) that there exists a constant $c_K > 0$ such that for all functions $\boldsymbol{\gamma}, \boldsymbol{\gamma}'$ on $\mathscr{W}$ with values in K,

$$h^2\left(R_{\boldsymbol{\gamma}}, R_{\boldsymbol{\gamma}'}\right) \geq c_K^2 \left\|\boldsymbol{\gamma} - \boldsymbol{\gamma}'\right\|_2^2.$$

Assumption 2.4.1 makes the Hellinger-type distance $h\left(R_{\boldsymbol{\gamma}}, R_{\boldsymbol{\gamma}'}\right)$ and the $\mathbb{L}_2(P_W)$-one between $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$ comparable, at least when $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$ take their values in K.

Proposition 2.4.3. *Let $\alpha \in (0,1]$ and $M > 0$. If P_W is the uniform distribution on $[0,1]$ and Assumption 2.4.1 is satisfied for a compact interval K of length $2\overline{L} > 0$, then*

$$\mathcal{R}_n(\mathcal{H}_\alpha(M)) \geq \frac{c_K^2}{48}\left[\left(\frac{3M^{1/\alpha}}{2^{2\alpha+4+1/\alpha}\kappa^2 n}\right)^{\frac{2\alpha}{1+2\alpha}} \bigwedge \left(\frac{M^2}{4}\right) \bigwedge \overline{L}^2\right].$$

This result says that in all exponential families for which Assumption 2.4.1 is satisfied, the order of magnitude of the minimax rate over $\mathcal{H}_\alpha(M)$ cannot be smaller than $n^{-2\alpha/(2\alpha+1)}$, at least when P_W is the uniform distribution on $[0,1]$.

2.4.2 A counterexample

Without a suitable parametrization of the exponential family $\overline{\mathscr{Q}} = \{Q_\theta,\ \theta \in I\}$ like that provided by Proposition 2.4.1, the minimax rate of convergence of $\mathcal{R}_n(\mathcal{H}_\alpha(M))$ may be different from $n^{-2\alpha/(2\alpha+1)}$ as shown by the following simple example of Poisson distributions parametrized by their means.

Proposition 2.4.4. *Let $\alpha \in (0,1]$, $M > 0$, P_W be the uniform distribution on $[0,1]$ and $\overline{\mathscr{Q}}$ the set of Poisson distributions R_γ with means $\gamma \in J = (0,+\infty)$. For all $n \geq 1$,*

$$\mathcal{R}_n(\mathcal{H}_\alpha(M)) \geq \frac{(1-e^{-1})}{144}\left[\left(\frac{3M^{1/\alpha}}{2^{4+\alpha+3/\alpha}n}\right)^{\frac{\alpha}{1+\alpha}} \bigwedge \frac{M}{8} \bigwedge \left(1+\frac{\sqrt{3}}{2}\right)\right].$$

In the Poisson case with this parametrization, the rate for $\mathcal{R}_n(\mathcal{H}_\alpha(M))$ is therefore at least of order $n^{-\alpha/(1+\alpha)}$, hence much slower than the one we would get if the family would be properly parametrized as indicated in the previous section, namely $n^{-2\alpha/(2\alpha+1)}$. We conclude that, depending on the exponential family, the parametrization by the mean may lead to different minimax rates. Nevertheless, as shown in the following proposition, the ρ-estimator still achieves the optimal rate (up to a logarithmic factor) in this case.

Proposition 2.4.5. *Let $\alpha \in (0,1]$, $M > 0$ and $\overline{\boldsymbol{S}}$ be the set of functions with values in $J = (0,+\infty)$ which are piecewise constant on each element of a partition $\{I_j, j \in \{1,\ldots,D\}\}$*

of $[0,1]$ *into* $D \geq 1$ *intervals of lengths* $1/D$. *For*

$$D = D(\alpha, M, n) = \min\left\{k \in \mathbb{N},\ \left(\frac{Mn}{2\log(en)}\right)^{\frac{1}{1+\alpha}} \leq k\right\},$$

the ρ*-estimator* $\widehat{\boldsymbol{\gamma}}$ *based on (any) countable and dense subset* $\boldsymbol{S}$ *of* $\overline{\boldsymbol{S}}$ *(with respect to supremum norm) satisfies*

$$\sup_{\boldsymbol{\gamma}^\star \in \mathcal{H}_\alpha(M)} \mathbb{E}\left[h^2\left(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}}\right)\right] \leq 2C'\left[\left(\frac{(M/2)^{1/\alpha}\log(en)}{n}\right)^{\frac{\alpha}{1+\alpha}} + \frac{3\log(en)}{2n}\right]$$

where C' *is the numerical constant appearing in (2.3.15).*

2.5 Calculation of ρ-estimators and simulation study

In this section, we study the performance of the ρ-estimator $\widehat{\boldsymbol{\theta}}$ of the regression function $\boldsymbol{\theta}^\star$ in the cases of Examples 2.2.2, 2.2.3, 2.2.4 which correspond respectively to the logit regression, Poisson and exponential distributions parametrized by their natural parameters.

The models

The function space $\overline{\boldsymbol{\Theta}}$ consists of functions $\boldsymbol{\theta}$ on $\mathscr{W} = \mathbb{R}^5$ with values in I and for $w = (w_1, \ldots, w_5) \in \mathscr{W}$ the value $\boldsymbol{\theta}(w)$ has the following form:
— In the Bernoulli model, $I = \mathbb{R}$ and

$$\boldsymbol{\theta}(w) = \eta_0 + \sum_{j=1}^{5} \eta_j w_j \quad \text{with } \eta = (\eta_0, \ldots, \eta_5) \in \mathbb{R}^6. \tag{2.5.1}$$

— In the Poisson model, $I = \mathbb{R}$ and

$$\boldsymbol{\theta}(w) = \log\log\left[1 + \exp\left(\eta_0 + \sum_{j=1}^{5} \eta_j w_j\right)\right] \quad \text{with } \eta = (\eta_0, \ldots, \eta_5) \in \mathbb{R}^6. \tag{2.5.2}$$

— In the exponential model, $I = (0, +\infty)$ and

$$\boldsymbol{\theta}(w) = \log\left[1 + \exp\left(\eta_0 + \sum_{j=1}^{5} \eta_j w_j\right)\right] \quad \text{with } \eta = (\eta_0, \ldots, \eta_5) \in \mathbb{R}^6. \tag{2.5.3}$$

For all these cases, the set $\overline{\boldsymbol{\Theta}}$ is VC-subgraph with dimension not larger than 7. For the calculation of the estimator on a computer, we do as if $\overline{\boldsymbol{\Theta}}$ were countable and consequently take $\boldsymbol{\Theta} = \overline{\boldsymbol{\Theta}}$.

The competitors

We compare the performance of $\widehat{\boldsymbol{\theta}}$ to that of the MLE and, in cases of Examples 2.2.3 and 2.2.4, to a median-based estimator $\widehat{\boldsymbol{\theta}}_0$. The estimator $\widehat{\boldsymbol{\theta}}_0$ is defined as any minimizer over $\overline{\boldsymbol{\Theta}}$ of the criterion

$$\boldsymbol{\theta} \mapsto \sum_{i=1}^{n} |Y_i - m(\boldsymbol{\theta}(W_i))| \tag{2.5.4}$$

where $m(\theta)$ is the median (or an approximation of it) of the distribution Q_θ for $\theta \in I$. We take $m(\theta) = e^\theta + 1/3 - 0.02e^{-\theta}$ for the Poisson distribution with parameter e^θ and $m(\theta) = (\log 2)/\theta$ for the exponential one with parameter θ.

In the examples we have chosen, the log-likelihood function is concave with respect to the parameter $\eta \in \mathbb{R}^6$ and the MLE is calculated using the stats4 R-package. The criterion (2.5.4) is not convex with respect to the parameter η and the median-based estimator is calculated using the cmaes R-package based on the CMA (Covariance Matrix Adaptation) method which turns out to be more stable than the gradient descent method. For more details about the CMA method, we refer the reader to Hansen (2016).

2.5.1 Calculation of the ρ-estimator

As mentioned in Section 2.3, we call ρ-estimator $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\boldsymbol{X})$ any element of the random set

$$\mathscr{E}(\boldsymbol{X}) = \left\{ \boldsymbol{\theta} \in \boldsymbol{\Theta} \text{ such that } \boldsymbol{\upsilon}(\boldsymbol{X}, \boldsymbol{\theta}) \leq \inf_{\boldsymbol{\theta}' \in \boldsymbol{\Theta}} \boldsymbol{\upsilon}(\boldsymbol{X}, \boldsymbol{\theta}') + \frac{\kappa_\rho}{25} \right\},$$

where

$$\boldsymbol{\upsilon}(\boldsymbol{X}, \boldsymbol{\theta}) = \sup_{\boldsymbol{\theta}' \in \boldsymbol{\Theta}} \mathbf{T}(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\theta}') = \sup_{\boldsymbol{\theta}' \in \boldsymbol{\Theta}} \sum_{i=1}^{n} \psi\left(\sqrt{\frac{q_{\boldsymbol{\theta}'}(X_i)}{q_{\boldsymbol{\theta}}(X_i)}} \right) \quad \text{for all } \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

To calculate the ρ-estimator $\widehat{\boldsymbol{\theta}}$ we use the iterative Algorithm 1 described below. We stop it either when the condition $\boldsymbol{\upsilon}(\boldsymbol{X}, \widehat{\boldsymbol{\theta}}) \leq 1$ is met or otherwise after $L = 100$ iterations. Since

$$\boldsymbol{\upsilon}(\boldsymbol{X}, \boldsymbol{\theta}) = \sup_{\boldsymbol{\theta}' \in \boldsymbol{\Theta}} \mathbf{T}(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\theta}') \geq \mathbf{T}(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\theta}) = 0,$$

the quantity $\inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \boldsymbol{\upsilon}(\boldsymbol{X}, \boldsymbol{\theta})$ is nonnegative and when $\boldsymbol{\upsilon}(\boldsymbol{X}, \widehat{\boldsymbol{\theta}}) \leq 1$,

$$\boldsymbol{\upsilon}(\boldsymbol{X}, \widehat{\boldsymbol{\theta}}) \leq \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \boldsymbol{\upsilon}(\boldsymbol{X}, \boldsymbol{\theta}) + 1 \leq \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \boldsymbol{\upsilon}(\boldsymbol{X}, \boldsymbol{\theta}) + \frac{\kappa_\rho}{25},$$

which shows that $\widehat{\boldsymbol{\theta}}$ is a ρ-estimator. The constant 1 has nothing magical, we just believe that the closer $\boldsymbol{\upsilon}(\boldsymbol{X}, \widehat{\boldsymbol{\theta}})$ to $\inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \boldsymbol{\upsilon}(\boldsymbol{X}, \boldsymbol{\theta})$ the better $\widehat{\boldsymbol{\theta}}$ performs. The condition $\boldsymbol{\upsilon}(\boldsymbol{X}, \widehat{\boldsymbol{\theta}}) \leq 1$ can therefore be seen as an early stopping time that guarantees that the estimator $\widehat{\boldsymbol{\theta}}$

almost minimizes $\boldsymbol{\theta} \mapsto \boldsymbol{\upsilon}(\boldsymbol{X}, \boldsymbol{\theta})$ over $\boldsymbol{\Theta}$. In all our simulations, (including the cases when we stop after 100 iterations), the resulting estimators $\widehat{\boldsymbol{\theta}}$ satisfy

$$\boldsymbol{\upsilon}(\boldsymbol{X}, \widehat{\boldsymbol{\theta}}) \leq \frac{\kappa_\rho}{25} \quad \text{hence} \quad \boldsymbol{\upsilon}(\boldsymbol{X}, \widehat{\boldsymbol{\theta}}) \leq \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \boldsymbol{\upsilon}(\boldsymbol{X}, \boldsymbol{\theta}) + \frac{\kappa_\rho}{25}$$

and are therefore ρ-estimators.

The algorithm is based on the following heuristic that we describe in the situation where the data are i.i.d. with distribution $P^\star = P_{\boldsymbol{\theta}^\star}$ for the sake of simplicity. It is proven in Proposition 14 of Baraud (2021), that $\mathbf{T}(\boldsymbol{X}, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$ is a good test statistic for testing $\mathcal{H}_0$: "$P_{\boldsymbol{\theta}_0}$ is closer (in Hellinger distance) to $P^\star$ than $P_{\boldsymbol{\theta}_1}$" against $\mathcal{H}_1$: "$P_{\boldsymbol{\theta}_1}$ is closer to $P^\star$ than $P_{\boldsymbol{\theta}_0}$". More precisely, with a probability close to 1, $\mathbf{T}(\boldsymbol{X}, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1) > 0$ when $h^2(P_{\boldsymbol{\theta}_1}, P^\star) \ll h^2(P_{\boldsymbol{\theta}_0}, P^\star)$ and $1/n \ll h^2(P_{\boldsymbol{\theta}_0}, P^\star)$ while the test statistic $\mathbf{T}(\boldsymbol{X}, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1) < 0$ when $h^2(P_{\boldsymbol{\theta}_0}, P^\star) \ll h^2(P_{\boldsymbol{\theta}_1}, P^\star)$ and $1/n \ll h^2(P_{\boldsymbol{\theta}_1}, P^\star)$. Note that if $h^2(P_{\boldsymbol{\theta}}, P^\star) \approx 1/n$ for both $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and $\boldsymbol{\theta} = \boldsymbol{\theta}_1$, the two distributions $P_{\boldsymbol{\theta}_0}$ and $P_{\boldsymbol{\theta}_1}$ are both close to $P^\star$ and choosing between $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ is unimportant. Because of these properties, if we start from an initial point $\boldsymbol{\theta}_0$ that is not too far from $\boldsymbol{\theta}^\star$ and if $\boldsymbol{\theta}_1$ is such that $\mathbf{T}(\boldsymbol{X}, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1) > 0$, it is likely that one of the two following situations occur:
— the quantity $h^2(P_{\boldsymbol{\theta}_1}, P^\star)$ is smaller or at least of comparable order as $h^2(P_{\boldsymbol{\theta}_0}, P^\star)$;
— $h^2(P_{\boldsymbol{\theta}_1}, P^\star)$ is of order $1/n$.
In any case, either $\boldsymbol{\theta}_1$ improves on $\boldsymbol{\theta}_0$ or, at least, performs similarly. We can then repeat the test starting now from $\boldsymbol{\theta}_1$ and looking from some $\boldsymbol{\theta}_2$ such that $\mathbf{T}(\boldsymbol{X}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) > 0$ and so on.

Since ρ-estimators are not unique, there is no reason for the algorithm to converge to a point and we are not expecting the algorithm to do so. Since the algorithm is based on the test statistic $\mathbf{T}(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\theta}')$, that provides a robust test between the probabilities $\mathbf{P}_{\boldsymbol{\theta}}$ and $\mathbf{P}_{\boldsymbol{\theta}'}$, as explained above, we expect the algorithm to get closer to the truth as we iterate it. As we shall see, only few iterations are in general necessary to meet the condition $\boldsymbol{\upsilon}(\boldsymbol{X}, \widehat{\boldsymbol{\theta}}) \leq 1$ and when it is not the case, the estimator obtained after $L = 100$ iterations provides a suitable estimation of the parameter. To find a maximizer of the mapping $\boldsymbol{\theta} \mapsto \mathbf{T}(\boldsymbol{X}, \widehat{\boldsymbol{\theta}}, \boldsymbol{\theta})$ at each iteration, we use the cmaes R-package.

To intialize the process we choose the value of $\boldsymbol{\theta}_0$ as follows. In the case of Bernoulli regression, we take for $\boldsymbol{\theta}_0$ the function on $\mathbb{R}^d$ that minimizes on $\overline{\boldsymbol{\Theta}}$ the penalized criterion (that can be found in the e1071 R-package)

$$\boldsymbol{\theta} \mapsto 10 \sum_{i=1}^{n} (1 - (2Y_i - 1)\boldsymbol{\theta}(W_i))_+ + \frac{1}{2} \sum_{i=1}^{d} |\boldsymbol{\theta}(e_i) - \boldsymbol{\theta}(0)|^2 ,$$

where $e_1, \ldots, e_d$ denotes the canonical basis of $\mathbb{R}^d$ (with $d = 6$). The e1071 R-package is used for the purpose of classifying the Y_i from the W_i. For the other exponential families we choose for $\boldsymbol{\theta}_0$ the median-based estimator $\widehat{\boldsymbol{\theta}}_0$.

Algorithm 1 Searching for the ρ-estimator

Input:

$\boldsymbol{X} = (X_1, \cdots, X_n)$: the data

$\boldsymbol{\theta}_0$: the starting point

Output: $\widehat{\boldsymbol{\theta}}$

Initialize $l = 0$, $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$;
while $\boldsymbol{v}(\boldsymbol{X}, \widehat{\boldsymbol{\theta}}) > 1$ and $l \leq L$ **do**
  $l \leftarrow l + 1$
  $\boldsymbol{\theta}_1 = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}} \mathbf{T}(\boldsymbol{X}, \widehat{\boldsymbol{\theta}}, \boldsymbol{\theta})$
  $\widehat{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}_1$
end while
Return $\widehat{\boldsymbol{\theta}}$.

2.5.2 When the model is exact

Throughout this section, we assume that the data $X_1, \ldots, X_n$ are i.i.d. with distribution $P_{\boldsymbol{\theta}^\star} = Q_{\boldsymbol{\theta}^\star} \cdot P_W$, $\boldsymbol{\theta}^\star \in \overline{\boldsymbol{\Theta}}$, and we estimate the risk

$$R_n(\widetilde{\boldsymbol{\theta}}) = \mathbb{E}\left[h^2\left(P_{\boldsymbol{\theta}^\star}, P_{\widetilde{\boldsymbol{\theta}}}\right)\right] = \mathbb{E}\left[\int_{\mathscr{W}} h^2\left(Q_{\boldsymbol{\theta}^\star(w)}, Q_{\widetilde{\boldsymbol{\theta}}(w)}\right) dP_W(w)\right]$$

of an estimator $\widetilde{\boldsymbol{\theta}}(\boldsymbol{X})$ by the Monte Carlo method on the basis of 500 replications. For this simulation study $n = 500$. We recall that, for a natural exponential family,

$$h^2(Q_\theta, Q_{\theta'}) = 1 - \exp\left[A\left(\frac{\theta + \theta'}{2}\right) - \frac{A(\theta) + A(\theta')}{2}\right] \tag{2.5.5}$$

where A is given in (2.2.1).

Bernoulli model. We consider the function $\boldsymbol{\theta}^\star = \boldsymbol{\theta}$ given by (2.5.1) with $\eta = (1, \ldots, 1) \in \mathbb{R}^6$. The distribution P_W is $(P_W^{(1)} + P_W^{(2)} + P_W^{(3)})/3$ where $P_W^{(1)}, P_W^{(2)}$ and $P_W^{(3)}$ are respectively the uniform distributions on the cubes

$$[-a, a]^5, \quad [b - 0.25, b + 0.25]^5 \quad \text{and} \quad [-b - 0.25, -b + 0.25]^5$$

with $a = 0.25$ and $b = 2$.

Poisson model In this case $\boldsymbol{\theta}^\star = \boldsymbol{\theta}$ given by (2.5.2) with $\eta = (0.7, 3, 4, 10, 2, 5)$. The distribution P_W is $P_{W,1}^{\otimes 2} \otimes P_{W,2} \otimes P_{W,3}^{\otimes 2}$ where $P_{W,1}, P_{W,2}$ and $P_{W,3}$ are the uniform distributions on $[0.2, 0.25]$, $[0.2, 0.3]$ and $[0.1, 0.2]$ respectively.

Exponential model We set $\boldsymbol{\theta}^\star = \boldsymbol{\theta}$ given by (2.5.3) with $\eta = (0.07, 3, 4, 6, 2, 1)$. The distribution P_W is $P_{W,1}^{\otimes 3} \otimes P_{W,2}^{\otimes 2}$ where $P_{W,1}$ and $P_{W,2}$ are the uniform distributions on $[0, 0.01]$ and $[0, 0.1]$ respectively.

In order to compare the performance of the ρ-estimator to the two other competitors we proceed as follows: we estimate the risk $R_n(\widehat{\boldsymbol{\theta}})$ of $\widehat{\boldsymbol{\theta}}$ by Monte Carlo as explained before. We then use this quantity as a benchmark and given another estimator $\widetilde{\boldsymbol{\theta}}$ we compute the quantity

$$\mathcal{E}(\widetilde{\boldsymbol{\theta}}) = \frac{R_n(\widetilde{\boldsymbol{\theta}}) - R_n(\widehat{\boldsymbol{\theta}})}{R_n(\widehat{\boldsymbol{\theta}})} \quad \text{so that} \quad R_n(\widetilde{\boldsymbol{\theta}}) = \Big(1 + \mathcal{E}(\widetilde{\boldsymbol{\theta}})\Big) R_n(\widehat{\boldsymbol{\theta}}). \tag{2.5.6}$$

Note that large positive values of $\mathcal{E}(\widetilde{\boldsymbol{\theta}})$ indicate a significant superiority of our estimator as compared to $\widetilde{\boldsymbol{\theta}}$ and negative values inferiority. The respective values of $R_n(\widehat{\boldsymbol{\theta}})$ and $\mathcal{E}(\widetilde{\boldsymbol{\theta}})$ are displayed in Table 2.1. The computation time of each estimator is displayed in Table 2.2.

Table 2.1: Values of $R_n(\widehat{\boldsymbol{\theta}})$ and $\mathcal{E}(\widetilde{\boldsymbol{\theta}})$ when the model is well-specified

	$R_n(\widehat{\boldsymbol{\theta}})$	$\mathcal{E}(\text{MLE})$	$\mathcal{E}(\widehat{\boldsymbol{\theta}}_0)$
Logit	0.0015	$< +0.1\%$	–
Poisson	0.0015	$< +0.1\%$	$+450\%$
Exponential	0.0015	$< +0.1\%$	$+110\%$

Table 2.2: Average computation time when the model is well-specified

	ρ-estimator	MLE	Median-based
Logit	331.43s	0.17s	–
Poisson	216.23s	0.23s	34.69s
Exponential	87.78s	0.28s	16.31s

Since the median of the Bernoulli distribution is either 0 or 1, hence only weakly depends on the value of the parameter, there is no estimator of the regression function based on the median for the Bernoulli model.

We observe the following facts:

- When the model is correct, the risks of the MLE and $\widehat{\boldsymbol{\theta}}$ are the same (the value of $\mathcal{E}(\text{MLE})$ is not larger than $1/1000$). In fact, a look at the simulations shows that the ρ-estimator coincides most of the time with the MLE, a fact which is consistent with the result proved in Section 5 of Baraud et al. (2017) that states the following:

under suitable (strong enough) assumptions, the MLE is a ρ-estimator when the statistical model is regular, exact and n is large enough. Our simulations indicate that the result actually holds under weaker assumptions.

- Both the MLE and the ρ-estimator outperform the median-based estimator $\widehat{\boldsymbol{\theta}}_0$.

- The quantities $R_n(\widehat{\boldsymbol{\theta}})$ are of order 0.0015 in all three cases. This fact can be explained as follows. In a regular statistical model $\mathscr{M}_0 = \{P_{\boldsymbol{\eta}},\ \boldsymbol{\eta} \in S\}$ parametrized with a parameter $\boldsymbol{\eta} \in S \subset \mathbb{R}^d$, the asymptotic normality properties of the MLE $\widehat{\boldsymbol{\eta}}_n$ together with the local equivalence of the Hellinger distance with the Euclidean one imply that, when the data are i.i.d. with distribution $P_{\boldsymbol{\eta}^\star} \in \mathscr{M}_0$,

$$n\mathbb{E}\left[h^2(P_{\widehat{\boldsymbol{\eta}}_n}, P_{\boldsymbol{\eta}^\star})\right] \underset{n\to+\infty}{\longrightarrow} \frac{d}{8}.$$

  In our simulation, conditionally to W, the distribution of Y is given by an exponential family parametrized by $d = 6$ parameters and the number of data being $n = 500$, we expect a risk of order $d/(8n) = 0.0015$, which is exactly what we obtained.

- The above result provides evidence that the algorithm we use does calculate the ρ-estimator as expected.

- In all the simulations we carried out, the algorithm required at most **two iterations** before the stopping condition $\boldsymbol{\upsilon}(\boldsymbol{X}, \widehat{\boldsymbol{\theta}}) \leq 1$ was met.

In the Bernoulli model, we also consider the case where the true regression function $\boldsymbol{\theta}^\star = \boldsymbol{\theta}$ is given by (2.5.1) with $\eta = (1, \ldots, 1) \in \mathbb{R}^6$ and $P_W = (P_W^{(2)} + P_W^{(3)})/2$. In such a situation, the MLE is likely not to exist because the sets of data labelled by 1 and 0 respectively can be perfectly separated by a hyperplane with probability close to 1. As expected the stats4 R-package for calculating the MLE returns an error. In contrast, the ρ-estimator always exists and its estimated risk $R_n(\widehat{\boldsymbol{\theta}})$ is of order 0.000179. In the 100 simulations we carried out, the algorithm stops after at most 2 iterations.

2.5.3 In presence of outliers

We now work with $n = 501$ independent random variables $X_1, \ldots, X_n$. The 500 first variables $X_1, \ldots, X_{n-1}$ are i.i.d. with distribution $P_{\boldsymbol{\theta}^\star}$ and simply follow the framework of the previous section. The last observation is chosen as follows. In the Bernoulli model $W_n = 1000(1, 1, 1, 1, 1)$ and $Y_n = -1$, for the Poisson case $W_n = 0.1(1, 1, 1, 1, 1)$ and $Y_n = 200$ and for the exponential case $W_n = 5 \times 10^{-3}(1, 1, 1, 10, 10)$ and $Y_n = 1000$. The results are displayed in Table 2.3 on the basis of 500 replications. The computation time for each estimator are given in Table 2.4.

We observe the following facts:

Table 2.3: Values of $R_n(\widehat{\boldsymbol{\theta}})$ and $\mathcal{E}(\widetilde{\boldsymbol{\theta}})$ in presence of an outlier

	$R_n(\widehat{\boldsymbol{\theta}})$	$\mathcal{E}(\text{MLE})$	$\mathcal{E}(\widehat{\boldsymbol{\theta}}_0)$
Logit	0.0015	+13000%	–
Poisson	0.0019	+1900%	+330%
Exponential	0.0018	+6000%	+78%

Table 2.4: Average computation time in presence of an outlier

	ρ-estimator	MLE	Median-based
Logit	497.31s	0.12s	–
Poisson	229.36s	0.29s	35.83s
Exponential	103.05s	0.32s	15.18s

- the risks of the ρ-estimator are quite similar to those given in Table 2.1 despite the presence of an outlier among the data set;
- the MLE behaves poorly;
- the performance of $\widehat{\boldsymbol{\theta}}$ remains much better than that of the median-based estimator $\widehat{\boldsymbol{\theta}}_0$.

Let us now display the quartiles of the distribution of the number of iterations that have been necessary to compute the ρ-estimator.

Table 2.5: Quartiles for the number of iterations in presence of outliers

	1st Quartile	Median	3rd Quartile	Maximum
Logit	3	3	3	6
Poisson	2	2	2	3
Exponential	2	2	2	3

Table 2.5 shows that the computation of the ρ-estimator requires only a few iterations of the algorithm.

2.5.4 When the data are contaminated

We now set $n = 500$ and define $P_{\boldsymbol{\theta}^\star}$ and P_W as in Section 2.5.2. We now assume that $X_1, \ldots, X_n$ are i.i.d. with distribution $P^\star = 0.95P_{\boldsymbol{\theta}^\star} + 0.05R$ for some (contaminating)

distribution R on $\mathscr{W} \times \mathscr{Y}$ with first marginal given by P_W. We restrict ourselves to the Poisson and exponential cases (we exclude the Bernoulli model since the Bernoulli distribution remains stable under the contamination by another Bernoulli distribution). In the Poisson case, we choose for R the distribution of the random variable $(W, 80 + B)$ where the conditional distribution of B given $W = (w_1, \ldots, w_5)$ is Bernoulli with mean $(1 + \exp[-(w_1 - w_2 - w_4 + w_5)])^{-1}$. In the case of the exponential distribution $R = P_W \otimes \mathcal{U}([50, 60])$ where $\mathcal{U}([50, 60])$ denotes the uniform distribution on $[50, 60]$.

We measure the performance of an estimator $\widetilde{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}^\star$ by means of the quantity

$$\overline{R}_n(\widetilde{\boldsymbol{\theta}}) = \mathbb{E}\left[h^2\left(P^\star, P_{\widetilde{\boldsymbol{\theta}}}\right)\right]$$

that we evaluate by Monte Carlo on the basis of 500 replications. We compare the performance of $\widehat{\boldsymbol{\theta}}$ to a competitor $\widetilde{\boldsymbol{\theta}}$ by evaluating the quantity

$$\overline{\mathcal{E}}(\widetilde{\boldsymbol{\theta}}) = \frac{\overline{R}_n(\widetilde{\boldsymbol{\theta}}) - \overline{R}_n(\widehat{\boldsymbol{\theta}})}{\overline{R}_n(\widehat{\boldsymbol{\theta}})}. \tag{2.5.7}$$

The results are displayed in Table 2.6 and the computation times in Table 2.7.

Table 2.6: Values of $\overline{R}_n(\widehat{\boldsymbol{\theta}})$ and $\overline{\mathcal{E}}(\widetilde{\boldsymbol{\theta}})$ under contamination (5%)

	$\overline{R}_n(\widehat{\boldsymbol{\theta}})$	$\overline{\mathcal{E}}$(MLE)	$\overline{\mathcal{E}}(\widehat{\boldsymbol{\theta}}_0)$
Poisson	0.028	+760%	+11%
Exponential	0.040	+320%	–17%

Table 2.7: Average computation time under contamination (5%)

	ρ-estimator	MLE	Median-based
Poisson	867.50s	0.33s	39.23s
Exponential	1863.97s	0.30s	20.47s

Let us now comment these results.

- With our choices of the contaminating distributions R, the (squared) Hellinger distance between the true distribution $P^\star$ of the data and the model is of order $h^2\left(P^\star, P_{\boldsymbol{\theta}^\star}\right) \approx 0.025$. As expected, we get that $\overline{R}_n(\widehat{\boldsymbol{\theta}}) \geq 0.025 \approx h^2\left(P^\star, P_{\boldsymbol{\theta}^\star}\right)$. Note that the situation is extreme in the sense that the approximation error is much larger than estimation error that can be achieved when the model is well specified (which is about 0.0015). This means that the model is "very" misspecified.
- The MLE behaves poorly.

- In the exponential case, the median-based estimator $\widehat{\boldsymbol{\theta}}_0$ outperforms the ρ-estimator while the opposite situation occurs in the Poisson case.

Table 2.8: Quartiles for the number of iterations when the data are contaminated

	1st Quartile	Median	3rd Quartile	Maximum
Poisson	5	5	5	100
Exponential	5	10	30	100

In Table 2.8, we observe that the number of iterations for calculating the ρ-estimator increases substantially as compared to the two previous situations. We note that for some simulations the algorithm was iterated 100 times (which corresponds to the maximal number of iterations that we allow) and the stopping condition $\boldsymbol{\upsilon}(\boldsymbol{X}, \widehat{\boldsymbol{\theta}}) \leq 1$ was not met (but satisfies $\boldsymbol{\upsilon}(\boldsymbol{X}, \widehat{\boldsymbol{\theta}}) \leq \kappa_\rho/25$). Despite this fact, the estimator that we get at the final step, hence after 100 iterations, performs well since the values of the risks $\overline{R}_n(\widehat{\boldsymbol{\theta}})$ are of the same order as $h^2\left(P^\star, P_{\boldsymbol{\theta}^\star}\right)$ and comparable to the median-based estimator $\widehat{\boldsymbol{\theta}}_0$.

2.6 Bounding the expectation of the supremum of an empirical process

The aim of this section is to prove the following result which will used later as an elementary material to prove Theorem 2.3.1.

Theorem 2.6.1. *Let $X_1, ..., X_n$ be n independent random variables with values in $(\mathscr{X}, \mathcal{X})$ and $\mathscr{F}$ an at most countable VC-subgraph class of functions with values in $[-1, 1]$ and VC-dimension not larger than $V \geq 1$. If*

$$Z(\mathscr{F}) = \sup_{f \in \mathscr{F}} \left| \sum_{i=1}^{n} (f(X_i) - \mathbb{E}\left[f(X_i)\right]) \right| \quad \textit{and} \quad \sup_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[f^2(X_i)\right] \leq \sigma^2 \leq 1,$$

then

$$\mathbb{E}\left[Z(\mathscr{F})\right] \leq 4.74\sqrt{nV\sigma^2\mathscr{L}(\sigma)} + 90V\mathscr{L}(\sigma), \tag{2.6.1}$$

with $\mathscr{L}(\sigma) = 9.11 + \log(1/\sigma^2)$.

Let us now turn to the proof. It follows from classical symmetrisation arguments that $\mathbb{E}\left[Z(\mathscr{F})\right] \leq 2\mathbb{E}\left[\overline{Z}(\mathscr{F})\right]$, where $\overline{Z}(\mathscr{F}) = \sup_{f \in \mathscr{F}} \left| \sum_{i=1}^{n} \varepsilon_i f(X_i) \right|$ and $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. Rademacher random variables. It is therefore enough to prove that

$$\mathbb{E}\left[\overline{Z}(\mathscr{F})\right] \leq 2.37\sqrt{nV\sigma^2\mathscr{L}(\sigma)} + 45V\mathscr{L}(\sigma). \tag{2.6.2}$$

Given a probability P and a class of functions $\mathscr{G}$ on $(E, \mathcal{E})$ we denote by $N_r(\epsilon, \mathscr{G}, P)$ the smallest cardinality of an ϵ-net for the $\mathbb{L}_r(E, \mathcal{E}, P)$-norm $\|\cdot\|_{r,P}$, i.e. the minimal cardinality of a subset $\mathscr{G}[\epsilon]$ of $\mathscr{G}$ that satisfies for all $g \in \mathscr{G}$

$$\inf_{\overline{g} \in \mathscr{G}[\epsilon]} \|g - \overline{g}\|_{r,P} = \inf_{\overline{g} \in \mathscr{G}[\epsilon]} \left(\int_E |g - \overline{g}|^r \, dP \right)^{1/r} \leq \epsilon.$$

We start with the following lemma.

Lemma 2.6.1. *Whatever the probability P on $(\mathscr{X}, \mathcal{X})$, $\epsilon \in (0, 2)$ and $r \geq 1$*

$$N_r(\epsilon, \mathscr{F}, P) \leq e(V+1)(2e)^V \left(\frac{2}{\epsilon} \right)^{rV}.$$

Proof. Let λ be the Lebesgue measure on $([-1,1], \mathscr{B}([-1,1]))$ and Q the product probability $P \otimes (\lambda/2)$ on $(E, \mathcal{E}) = (\mathscr{X} \times [-1,1], \mathcal{X} \times \mathscr{B}([-1,1]))$. Given two elements $f, g \in \mathscr{F}$ and $x \in \mathscr{X}$

$$\begin{aligned} \int_{[-1,1]} \left| \mathbb{1}_{f(x)>t} - \mathbb{1}_{g(x)>t} \right| dt &= \int_{[-1,1]} \left(\mathbb{1}_{f(x)>t\geq g(x)} + \mathbb{1}_{g(x)>t\geq f(x)} \right) dt \\ &= |f(x) - g(x)| \end{aligned}$$

and, setting $C_f = \{(x,t) \in \mathscr{X} \times [-1,1],\ f(x) > t\}$ the subgraph of f and similarly C_g that of g, we deduce from Fubini's theorem that

$$\begin{aligned} \|f - g\|_{1,P} = \int_{\mathscr{X}} |f - g| \, dP &= 2 \int_{\mathscr{X} \times [-1,1]} \left| \mathbb{1}_{C_f}(x,t) - \mathbb{1}_{C_g}(x,t) \right| dQ \\ &= 2 \left\| \mathbb{1}_{C_f} - \mathbb{1}_{C_g} \right\|_{1,Q}. \end{aligned}$$

Since the functions $f, g \in \mathscr{F}$ take their values in $[-1,1]$,

$$\|f - g\|_{r,P}^r = \int_{\mathscr{X}} |f - g|^r \, dP \leq 2^{r-1} \int_{\mathscr{X}} |f - g| \, dP \leq 2^r \left\| \mathbb{1}_{C_f} - \mathbb{1}_{C_g} \right\|_{1,Q}$$

and consequently, for all $\epsilon > 0$

$$N_r(\epsilon, \mathscr{F}, P) \leq N_1((\epsilon/2)^r, \mathscr{G}, Q) \quad \text{with} \quad \mathscr{G} = \{\mathbb{1}_{C_f},\ f \in \mathscr{F}\}.$$

Since $\mathscr{F}$ is VC-subgraph with VC-dimension not larger than V, the class $\mathscr{G}$ is by definition VC with dimension not larger than V and the result follows from Corollary 1 in Haussler (1995). □

The proof of Theorem 2.6.1 is based on a chaining argument. It follows from the monotone convergence theorem that it is actually enough to prove (2.6.2) with $\mathscr{F}_J$, $J \geq 1$, in place of $\mathscr{F}$ where $(\mathscr{F}_J)_{J \geq 1}$ is a sequence of finite subsets of $\mathscr{F}$ which is increasing for the inclusion and satisfies $\bigcup_{J \geq 1} \mathscr{F}_J = \mathscr{F}$. We may therefore assume with no loss of generality that $\mathscr{F}$ is finite.

Let q be some positive number in $(0,1)$ to be chosen later on and $P_{\boldsymbol{X}}$ the empirical distribution $n^{-1}\sum_{i=1}^{n}\delta_{X_i}$. We shall denote by $\mathbb{E}_\varepsilon$ the expectation with respect to the Rademacher random variables ε_i, hence conditionally on $\boldsymbol{X}=(X_1,\ldots,X_n)$. Let $\|\cdot\|_{2,\boldsymbol{X}}$ be the $\mathbb{L}_2(\mathscr{X},\mathcal{X},P_{\boldsymbol{X}})$-norm and

$$\widehat{\sigma}^2=\widehat{\sigma}^2(\boldsymbol{X})=\sup_{f\in\mathscr{F}}\|f\|_{2,\boldsymbol{X}}^2=\sup_{f\in\mathscr{F}}\left[\frac{1}{n}\sum_{i=1}^{n}f^2(X_i)\right]\in[0,1].$$

For each positive integer k, let $\mathscr{F}_k=\mathscr{F}_k(\boldsymbol{X})$ be a minimal $(q^k\widehat{\sigma})$-net for $\mathscr{F}$ with respect to $\|\cdot\|_{2,\boldsymbol{X}}$. In particular, we can associate to a function $f\in\mathscr{F}$ a sequence $(f_k)_{k\geq 1}$ with $f_k\in\mathscr{F}_k$ satisfying $\|f-f_k\|_{2,\boldsymbol{X}}\leq q^k\widehat{\sigma}$ for all $k\geq 1$. Actually, since $\mathscr{F}$ is finite $f_k=f$ for all k large enough. Besides, it follows from Lemma 2.6.1 with the choices $r=2$ and $P=P_{\boldsymbol{X}}$ that for all $k\geq 1$ we can choose $\mathscr{F}_k$ in such a way that $\log[\operatorname{Card}\mathscr{F}_k]$ is not larger than $h(q^k\widehat{\sigma})$ where

$$h(\epsilon)=\log\left[e(V+1)(2e)^V\right]+2V\log\left(\frac{2}{\epsilon}\right)\quad\text{for all }\epsilon\in(0,1]. \tag{2.6.3}$$

For $f\in\mathscr{F}$, the following (finite) decomposition holds

$$\begin{aligned}\sum_{i=1}^{n}\varepsilon_i f(X_i)&=\sum_{i=1}^{n}\varepsilon_i f_1(X_i)+\sum_{i=1}^{n}\varepsilon_i\sum_{k=1}^{+\infty}[f_{k+1}(X_i)-f_k(X_i)]\\&=\sum_{i=1}^{n}\varepsilon_i f_1(X_i)+\sum_{k=1}^{+\infty}\left[\sum_{i=1}^{n}\varepsilon_i\left(f_{k+1}(X_i)-f_k(X_i)\right)\right].\end{aligned}$$

Setting $\mathscr{F}_k^2=\{(f_k,f_{k+1}),\ f\in\mathscr{F}\}$ for all $k\geq 1$, we deduce that

$$\overline{Z}(\mathscr{F})\leq\sup_{f\in\mathscr{F}_1}\left|\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|+\sum_{k=1}^{+\infty}\sup_{(f_k,f_{k+1})\in\mathscr{F}_k^2}\left|\sum_{i=1}^{n}\varepsilon_i\left[f_{k+1}(X_i)-f_k(X_i)\right]\right|$$

and consequently,

$$\begin{aligned}\mathbb{E}_\varepsilon\left[\overline{Z}(\mathscr{F})\right]\leq\mathbb{E}_\varepsilon&\left[\sup_{f\in\mathscr{F}_1}\left|\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|\right]\\&+\sum_{k=1}^{+\infty}\mathbb{E}_\varepsilon\left[\sup_{(f_k,f_{k+1})\in\mathscr{F}_k^2}\left|\sum_{i=1}^{n}\varepsilon_i\left[f_k(X_i)-f_{k+1}(X_i)\right]\right|\right].\end{aligned}$$

Given a finite set $\mathscr{G}$ of functions on $\mathscr{X}$ and setting $-\mathscr{G}=\{-g,\ g\in\mathscr{G}\}$ and $v^2=\max_{g\in\mathscr{G}}\|g\|_{2,\boldsymbol{X}}^2$, we shall repeatedly use the inequality

$$\mathbb{E}\left[\sup_{g\in\mathscr{G}}\left|\sum_{i=1}^{n}\varepsilon_i g(X_i)\right|\right]=\mathbb{E}\left[\sup_{g\in\mathscr{G}\cup(-\mathscr{G})}\sum_{i=1}^{n}\varepsilon_i g(X_i)\right]\leq\sqrt{2n\log(2\operatorname{Card}\mathscr{G})v^2}$$

that can be found in Massart (2007) (see inequality (6.3)). Since $\max_{f\in\mathscr{F}_1}\|f\|_{2,\boldsymbol{X}}^2 \le \widehat{\sigma}^2$, $\log(\operatorname{Card}\mathscr{F}_1) \le h(q\widehat{\sigma})$, $\log(\operatorname{Card}\mathscr{F}_k^2) \le h(q^k\widehat{\sigma}) + h(q^{k+1}\widehat{\sigma})$ and

$$\begin{aligned}&\sup_{(f_k,f_{k+1})\in\mathscr{F}_k^2}\|f_k - f_{k+1}\|_{2,\boldsymbol{X}}^2\\ &\quad\le \sup_{f\in\mathscr{F}}\sup_{(f_k,f_{k+1})\in\mathscr{F}_k^2}\left(\|f-f_k\|_{2,\boldsymbol{X}} + \|f-f_{k+1}\|_{2,\boldsymbol{X}}\right)^2 \le (1+q)^2 q^{2k}\widehat{\sigma}^2,\end{aligned}$$

we deduce that

$$\mathbb{E}_\varepsilon\left[\sup_{f\in\mathscr{F}_1}\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right] \le \widehat{\sigma}\sqrt{2n\left(\log 2 + h(q\widehat{\sigma})\right)},$$

and for all $k \ge 1$

$$\begin{aligned}&\mathbb{E}_\varepsilon\left[\sup_{(f,g)\in\mathscr{F}_k^2}\left|\sum_{i=1}^n \varepsilon_i\left[g(X_i) - f(X_i)\right]\right|\right]\\ &\qquad\qquad\le \widehat{\sigma}(1+q)q^k\sqrt{2n\left(\log 2 + h(q^k\widehat{\sigma}) + h(q^{k+1}\widehat{\sigma})\right)}.\end{aligned}$$

Setting $g : u \mapsto \sqrt{\log 2 + h(u) + h(qu)}$ on $(0,1]$ and using the fact that g is decreasing (since h is) we deduce that

$$\begin{aligned}&\mathbb{E}_\varepsilon\left[\overline{Z}(\mathscr{F})\right]\\ &\le \widehat{\sigma}\sqrt{2n}\left[\sqrt{\log 2 + h(q\widehat{\sigma})} + (1+q)\sum_{k\ge1} q^k\sqrt{\log 2 + h(q^k\widehat{\sigma}) + h(q^{k+1}\widehat{\sigma})}\right]\\ &\le \widehat{\sigma}\sqrt{2n}\left[g(\widehat{\sigma}) + (1+q)\sum_{k\ge1} q^k g(q^k\widehat{\sigma})\right]\\ &\le \sqrt{2n}\left[\frac{1}{1-q}\int_{q\widehat{\sigma}}^{\widehat{\sigma}} g(u)du + \frac{1+q}{1-q}\sum_{k\ge1}\int_{q^{k+1}\widehat{\sigma}}^{q^k\widehat{\sigma}} g(u)du\right]\\ &\le \sqrt{2n}\frac{1+q}{1-q}\int_0^{\widehat{\sigma}} g(u)du.\end{aligned}$$

The mapping g being positive and decreasing, the function $G : y \mapsto \int_0^y g(u)du$ is increasing and concave. Taking the expectation with respect to $\boldsymbol{X}$ on both sides of the previous inequality and using Jensen's inequality we get

$$\begin{aligned}\mathbb{E}\left[\overline{Z}(\mathscr{F})\right] &\le \sqrt{2n}\frac{1+q}{1-q}\mathbb{E}\left[G(\widehat{\sigma})\right] \le \sqrt{2n}\frac{1+q}{1-q}G\left(\mathbb{E}\left[\widehat{\sigma}\right]\right)\\ &\le \sqrt{2n}\frac{1+q}{1-q}G\left(\sqrt{\mathbb{E}\left[\widehat{\sigma}^2\right]}\right).\end{aligned} \tag{2.6.4}$$

By symmetrization and contraction arguments (see Theorem 4.12 in Ledoux and Talagrand (1991)),

$$\begin{aligned}\mathbb{E}\left[n\widehat{\sigma}^2\right] &\le \mathbb{E}\left[\sup_{f\in\mathscr{F}}\sum_{i=1}^n\left(f^2(X_i)-\mathbb{E}\left[f^2(X_i)\right]\right)\right]+\sup_{f\in\mathscr{F}}\sum_{i=1}^n\mathbb{E}\left[f^2(X_i)\right]\\ &\le 2\mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n\varepsilon_i f^2(X_i)\right|\right]+n\sigma^2\\ &\le 8\mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n\varepsilon_i f(X_i)\right|\right]+n\sigma^2=8\mathbb{E}\left[\overline{Z}(\mathscr{F})\right]+n\sigma^2 \qquad (2.6.5)\end{aligned}$$

and we infer from (2.6.4) that

$$\mathbb{E}\left[\overline{Z}(\mathscr{F})\right]\le\sqrt{2n}\frac{1+q}{1-q}G(B)\quad\text{with}\quad B=\sqrt{\sigma^2+\frac{8\mathbb{E}\left[\overline{Z}(\mathscr{F})\right]}{n}}\wedge 1. \qquad (2.6.6)$$

The following lemma provides an evaluation of G.

Lemma 2.6.2. *Let a, b, y_0 be positive numbers and $y\in[y_0,1]$,*

$$\int_0^y\sqrt{a+b\log(1/u)}du\le\left(1+\frac{b}{2a}\right)y\sqrt{a+b\log(1/y_0)}.$$

Proof. Using an integration by parts and the fact that

$$\frac{d}{du}\sqrt{a+b\log(1/u)}=-\frac{b}{2u\sqrt{a+b\log(1/u)}}$$

we get

$$\begin{aligned}\int_0^y\sqrt{a+b\log(1/u)}du&=\left[u\sqrt{a+b\log(1/u)}\right]_0^y+\frac{1}{2}\int_0^y\frac{b}{\sqrt{a+b\log(1/u)}}du\\ &\le y\sqrt{a+b\log(1/y)}+\frac{by}{2\sqrt{a+b\log(1/y)}}\\ &=y\sqrt{a+b\log(1/y)}\left[1+\frac{b}{2\left(a+b\log(1/y)\right)}\right]\end{aligned}$$

and the conclusion follows from the fact that $y_0\le y\le 1$. □

Since for all $y\in(0,1]$, $g(y)=\sqrt{a+b\log(1/y)}$ with

$$a=\log[2e^2(V+1)^2]+2V\log(8e/q)\quad\text{and}\quad b=4V$$

we may apply Lemma 2.6.2 with $y_0=\sigma$ and $y=B$ and deduce from (2.6.6) that

$$\begin{aligned}\mathbb{E}\left[\overline{Z}(\mathscr{F})\right]&\le\sqrt{2n}\frac{1+q}{1-q}\left(1+\frac{b}{2a}\right)B\sqrt{a+b\log(1/\sigma)}\\ &\le\sqrt{2n}\frac{1+q}{1-q}\left(1+\frac{b}{2a}\right)\sqrt{\sigma^2+\frac{8\mathbb{E}\left[\overline{Z}(\mathscr{F})\right]}{n}}\sqrt{a+b\log(1/\sigma)}.\end{aligned}$$

Solving the inequality $\mathbb{E}\left[\overline{Z}(\mathscr{F})\right] \leq A\sqrt{2n\sigma^2 + 16\mathbb{E}\left[\overline{Z}(\mathscr{F})\right]}$ with

$$A = \frac{1+q}{1-q}\left(1 + \frac{b}{2a}\right)\sqrt{a + b\log(1/\sigma)},$$

we get that

$$\mathbb{E}\left[\overline{Z}(\mathscr{F})\right] \leq 8A^2 + \sqrt{64A^4 + 2A^2 n\sigma^2} \leq 16A^2 + A\sqrt{2n\sigma^2}. \tag{2.6.7}$$

Finally, we conclude by using the inequalities

$$\begin{aligned}
\frac{b}{2a} &= \frac{4V}{2\left[\log[2e^2(V+1)^2] + 2V\log(8e/q)\right]} \leq \frac{1}{\log(8e/q)}, \\
\frac{a}{b} &= \frac{\log[2e^2(V+1)^2] + 2V\log(8e/q)}{4V} \\
&= \frac{\log(8e/q)}{2} + \frac{\log[2e^2(V+1)^2]}{4V} \leq \frac{\log(8e/q)}{2} + \frac{\log[8e^2]}{4} \\
&= \log\left(\frac{8^{3/4}e}{\sqrt{q}}\right)
\end{aligned}$$

which, with our choice $q = 0.0185$, give

$$\begin{aligned}
A &\leq \frac{1+q}{1-q}\left(1 + \frac{1}{\log(8e/q)}\right)\sqrt{4V\left(\log\left(\frac{8^{3/4}e}{\sqrt{q}}\right) + \log\frac{1}{\sigma}\right)} \\
&\leq 2.37\sqrt{V\left(4.555 + \log\frac{1}{\sigma}\right)}
\end{aligned}$$

and together with (2.6.7) leads to (2.6.2).

2.7 Proofs of main theorem and properties

2.7.1 Proof of Theorem 2.3.1

We recall that the function ψ defined by (2.3.1) satisfies Assumption 2 of Baraud and Birgé (2018) with $a_0 = 4$, $a_1 = 3/8$ and $a_2^2 = 3\sqrt{2}$ (see their Proposition 3). Theorem 2.3.1 is actually a consequence of Theorem 1 of Baraud and Birgé (2018). Set $\boldsymbol{\mu} = \bigotimes_{i=1}^n \mu_i$ with $\mu_i = P_{W_i} \otimes \nu$ for all $i \in \{1, \ldots, n\}$, denote by $\boldsymbol{\mathcal{Q}}$ the following families of densities (with respect to $\boldsymbol{\mu}$) on $\mathscr{X}^n = (\mathscr{W} \times \mathscr{Y})^n$

$$\boldsymbol{\mathcal{Q}} = \{\mathbf{p}_{\boldsymbol{\theta}} : \mathbf{x} = (x_1, \ldots, x_n) \mapsto q_{\boldsymbol{\theta}}(x_1)\ldots q_{\boldsymbol{\theta}}(x_n),\ \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$$

and by $\boldsymbol{\mathscr{P}}$ the corresponding ρ-model, i.e. the countable set $\{\mathbf{P} = \mathbf{p}_{\boldsymbol{\theta}} \cdot \boldsymbol{\mu},\ \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ with representation $(\boldsymbol{\mu}, \boldsymbol{\mathcal{Q}})$. We first prove the following results.

Proposition 2.7.1. *Under Assumption 2.3.1, the class of functions* $\mathcal{Q} = \{q_{\boldsymbol{\theta}} : (w, y) \mapsto q_{\boldsymbol{\theta}(w)}(y),\ \boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}}\}$ *on* $\mathscr{X} = \mathscr{W} \times \mathscr{Y}$ *is VC-subgraph with dimension not larger than* $9.41V$.

Proof. Since the exponential function is monotone on $\mathbb{R}$, by Proposition 1.5.2, it suffices to prove that the family

$$\mathscr{F} = \left\{ f : (w, y) \mapsto S(y)\boldsymbol{\theta}(w) - A(\boldsymbol{\theta}(w)),\ \boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}} \right\}$$

is VC-subgraph on $\mathscr{X} = \mathscr{W} \times \mathscr{Y}$ with dimension not larger than $9.41V$. The function A being convex and continuous on I, the mapping defined on I by $\theta \mapsto S(y)\theta - A(\theta)$ is continuous and concave for all fixed $y \in \mathscr{Y}$. In particular, for $u \in \mathbb{R}$ the level set $\{\theta \in I,\ S(y)\theta - A(\theta) > u\}$ is an open subinterval of I of the form $(\underline{a}(y,u), \overline{a}(y,u))$ where $\underline{a}(y,u)$ and $\overline{a}(y,u)$ belong to the closure $\overline{I}$ of I in $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$. For $\boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}}$, let us set

$$\begin{aligned} C^+_{\boldsymbol{\theta}} &= \{(w, b, b') \in \mathscr{W} \times \overline{I}^2,\ \boldsymbol{\theta}(w) > b\} \\ C^-_{\boldsymbol{\theta}} &= \{(w, b, b') \in \mathscr{W} \times \overline{I}^2,\ \boldsymbol{\theta}(w) < b'\} \end{aligned}$$

and define $\mathscr{C}^+$ (respectively $\mathscr{C}^-$) as the class of all subsets $C^+_{\boldsymbol{\theta}}$ (respectively $C^-_{\boldsymbol{\theta}}$) when $\boldsymbol{\theta}$ varies among $\overline{\boldsymbol{\Theta}}$.

Let us prove that $\mathscr{C}^+$ is a VC-class of sets on $\mathscr{Z} = \mathscr{W} \times \overline{I}^2$ with dimension not larger than V. If $\mathscr{C}^+$ shatters the finite subset $\{z_1, \ldots, z_k\}$ of $\mathscr{Z}$ with $z_i = (w_i, b_i, b'_i)$ for $i \in \{1, \ldots, k\}$, necessarily the b_i belong to $\mathbb{R}$ for all $i \in \{1, \ldots, k\}$.

Consequently, the class of subgraphs

$$\widetilde{\mathscr{C}}^+ = \left\{ \{(w, b) \in \mathscr{W} \times \mathbb{R},\ \boldsymbol{\theta}(w) > b\},\ \boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}} \right\}$$

shatters the points $\widetilde{z}_1 = (w_1, b_1), \ldots, \widetilde{z}_k = (w_k, b_k)$ in $\mathscr{W} \times \mathbb{R}$. This is possible only for $k \leq V$ since, by Assumption 2.3.1, $\overline{\boldsymbol{\Theta}}$ is VC-subgraph on $\mathscr{W}$ with dimension V.

Arguing similarly we obtain that $\mathscr{C}^-$ is also VC on $\mathscr{Z}$ with dimension not larger than V. In particular, it follows from Theorem 1.5.1 that the class of subsets

$$\mathscr{C}^+ \bigwedge \mathscr{C}^- = \{C^+ \cap C^-,\ \ C^+ \in \mathscr{C}^+,\ C^- \in \mathscr{C}^-\}$$

is VC on $\mathscr{Z}$ with dimension not larger than $9.41V$.

Let us now conclude the proof. If the class of subgraphs of $\mathscr{F}$ shatter the points $(w_1, y_1, u_1), \ldots, (w_k, y_k, u_k)$ in $\mathscr{W} \times \mathscr{Y} \times \mathbb{R}$, this means that for all subsets J of $\{1, \ldots, k\}$, there exists a function $\boldsymbol{\theta} = \boldsymbol{\theta}(J) \in \overline{\boldsymbol{\Theta}}$ such that the condition $j \in J$ is equivalent to the following ones

$$S(y_j)\boldsymbol{\theta}(w_j) - A\left(\boldsymbol{\theta}(w_j)\right) > u_j \iff \boldsymbol{\theta}(w_j) \in (\underline{a}(y_j, u_j), \overline{a}(y_j, u_j))$$

and finally equivalent to

$$z_j = (w_j, \underline{a}(y_j, u_j), \overline{a}(y_j, u_j)) \in C^+_{\boldsymbol{\theta}} \cap C^-_{\boldsymbol{\theta}}.$$

Hence, the class

$$\mathscr{C} = \left\{ C^+_{\boldsymbol{\theta}} \cap C^-_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}} \right\} \subset \mathscr{C}^+ \bigwedge \mathscr{C}^-$$

shatters $\{z_1, \ldots, z_k\}$ in $\mathscr{Z}$. This is possible for $k \leq 9.41V$ only and proves the fact that $\mathscr{F}$ is VC-subgraph with dimension not larger than $9.41V$. □

The result below provides an upper bound on the ρ-dimension function $(\mathbf{P}^\star, \overline{\mathbf{P}}) \mapsto D^{\mathscr{P}}(\mathbf{P}^\star, \overline{\mathbf{P}})$ of $\mathscr{P}$. The ρ-dimension function is defined by Definition 4 of Baraud and Birgé (2018).

Proposition 2.7.2. *Under Assumption 2.3.1, for all product probabilities* $\mathbf{P}^\star, \overline{\mathbf{P}} = \otimes_{i=1}^n \overline{P}_i$ *on* $(\mathscr{X}^n, \mathcal{X}^{\otimes n})$ *with* $\overline{P}_i = \overline{p} \cdot \mu_i$ *for all* $i \in \{1, \dots, n\}$,

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

Proof. Given two product probabilities $\mathbf{R} = \otimes_{i=1}^n R_i$ and $\mathbf{R}' = \otimes_{i=1}^n R_i'$ on $(\mathscr{X}^n, \mathcal{X}^{\otimes n})$, we set $\mathbf{h}^2(\mathbf{R}, \mathbf{R}') = \sum_{i=1}^n h^2(R_i, R_i')$ and for $y > 0$,

$$\mathscr{F}_y = \left\{ \psi\left(\sqrt{\frac{q_{\boldsymbol{\theta}}}{\overline{p}}}\right) \,\middle|\, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \mathbf{h}^2(\mathbf{P}^\star, \overline{\mathbf{P}}) + \mathbf{h}^2(\mathbf{P}^\star, \mathbf{p}_{\boldsymbol{\theta}} \cdot \boldsymbol{\mu}) < y^2 \right\}.$$

It follows from Proposition 2.7.1 and Proposition 1.5.2 that $\mathscr{F}_y$ is VC-subgraph with dimension not larger than $\overline{V} = 9.41V$. Besides, by Proposition 3 in Baraud and Birgé (2018) we know that our function ψ satisfies their Assumption 2 and more precisely (11) which, together with the definition of $\mathscr{F}_y$, implies that $\sup_{f \in \mathscr{F}_y} n^{-1} \sum_{i=1}^n \mathbb{E}\left[f^2(X_i)\right] \leq \sigma^2(y) = (a_2^2 y^2/n) \wedge 1$. Applying Theorem 2.6.1 with $\mathscr{F} = \mathscr{F}_y$, we obtain that

$$\begin{aligned} w^{\mathscr{P}}(\mathbf{P}^\star, \overline{\mathbf{P}}, y) &= \mathbb{E}\left[\sup_{f \in \mathcal{F}_y} \left|\sum_{i=1}^n f(X_i) - \mathbb{E}\left[f(X_i)\right]\right|\right] \\ &\leq 4.74 a_2 y \sqrt{\overline{V}\mathscr{L}(\sigma(y))} + 90\overline{V}\mathscr{L}(\sigma(y)) \\ &= 14.55 a_2 y \sqrt{V\mathscr{L}(\sigma(y))} + 846.9 V \mathscr{L}(\sigma(y)). \end{aligned}$$

Let $D \geq a_1^2 V/(16 a_2^4) = 2^{-11} V$ to be chosen later on and $\beta = a_1/(4a_2)$. For $y \geq \beta^{-1}\sqrt{D}$,

$$\begin{aligned} \mathscr{L}(\sigma(y)) &= 9.11 + \log_+\left(\frac{n}{a_2^2 y^2}\right) \leq 9.11 + \log_+\left(\frac{n}{a_2^2 \beta^{-2} D}\right) \\ &= 9.11 + \log_+\left(\frac{a_1^2 n}{16 a_2^4 D}\right) \leq 9.11 + \log_+\left(\frac{n}{V}\right) = L. \end{aligned}$$

Hence for all $y \geq \beta^{-1}\sqrt{D}$,

$$\begin{aligned} w^{\mathscr{P}}(\mathbf{P}^\star, \overline{\mathbf{P}}, y) &\leq 14.55 a_2 y \sqrt{VL} + 846.9 VL \\ &= \frac{a_1 y^2}{8}\left[\frac{8 \times 14.55 a_2 \sqrt{VL}}{a_1 y} + \frac{8 \times 846.9 VL}{a_1 y^2}\right] \\ &\leq \frac{a_1 y^2}{8}\left[\frac{8 \times 14.55 a_2 \sqrt{VL}}{a_1 \beta^{-1}\sqrt{D}} + \frac{8 \times 846.9 VL}{a_1 \beta^{-2} D}\right] \\ &= \frac{a_1 y^2}{8}\left[\frac{2 \times 14.55 \sqrt{VL}}{\sqrt{D}} + \frac{8 \times 846.9 a_1 VL}{16 a_2^2 D}\right] \\ &= \frac{a_1 y^2}{8}\left[\frac{29.1\sqrt{VL}}{\sqrt{D}} + \frac{37.5 VL}{D}\right] \leq \frac{a_1 y^2}{8} \end{aligned}$$

for $D = 10^3 VL > 2^{-11}V$. The result follows from the definition of the ρ-dimension. □

Let us now complete the proof of Theorem 2.3.1. It follows from Theorem 1 of Baraud and Birgé (2018) that the ρ-estimator $\widehat{\mathbf{P}} = \mathbf{P}_{\widehat{\boldsymbol{\theta}}}$ built on the ρ-model $\mathscr{P}$, which coincides with that described in Section 2.3.1, satisfies for all $\overline{\mathbf{P}} \in \mathscr{P}$, with a probability at least $1 - e^{-\xi}$,

$$\mathbf{h}^2(\mathbf{P}^\star, \widehat{\mathbf{P}}) \leq \gamma \mathbf{h}^2\left(\mathbf{P}^\star, \mathscr{P}\right) + \gamma' \left(\frac{D^{\mathscr{P}}(\mathbf{P}^\star, \overline{\mathbf{P}})}{4.7} + 1.49 + \xi \right)$$

with

$$\gamma = \frac{4(a_0+8)}{a_1} + 2 + \frac{84}{a_2^2} < 150 \quad \text{and} \quad \gamma' = \frac{4}{a_1}\left(\frac{35a_2^2}{a_1} + 74\right) < 5014$$

and $D^{\mathscr{P}}(\mathbf{P}^\star, \overline{\mathbf{P}}) \leq 10^3 V \left[9.11 + \log_+(n/V)\right]$ by Proposition 2.7.2. Finally, the result follows from the facts that $\mathbf{h}^2(\mathbf{P}^\star, \widehat{\mathbf{P}}) = \mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_{\widehat{\boldsymbol{\theta}}})$ and $\mathbf{h}^2\left(\mathbf{P}^\star, \mathscr{P}\right) = \mathbf{h}^2\left(\mathbf{Q}^\star, \mathscr{Q}\right)$.

2.7.2 A preliminary result

Proposition 2.7.3. *Let g be a 1-Lipschitz function on $\mathbb{R}$ supported on $[0,1]$, N some positive integer and L some positive number. For $\varepsilon \in \{-1,1\}^{2^N}$ define the function G_ε as*

$$G_\varepsilon(x) = L \sum_{k=0}^{2^N-1} \varepsilon_{k+1} g\left(2^N x - k\right) \quad \text{for all } x \in [0,1]. \tag{2.7.1}$$

Then, G_ε satisfies (2.4.1) with $\alpha \in (0,1]$ and $M > 0$ provided that $L \leq 2^{-[(N-1)\alpha+1]} M$.

Proof. For $k \in \Lambda = \{0, \ldots, 2^N - 1\}$, we set $g_k : x \mapsto g(2^N x - k)$. Since g is 1-Lipschitz and supported on $[0,1]$, the function g_k is 2^N-Lipschitz on $\mathbb{R}$ and supported on $I_k = [2^{-N}k, 2^{-N}(k+1)] \subset [0,1]$ for all $k \in \Lambda$. In particular, the intersection of the supports of g_k and $g_{k'}$ reduces to at most a singleton when $k \neq k'$.

Let $x < y$ be two points in $[0,1]$. If there exists $k \in \Lambda$ such that $x, y \in I_k$, using that $0 \leq y - x \leq 2^{-N}$ and the fact that $L2^{N\alpha} \leq L2^{(N-1)\alpha+1} \leq M$, we obtain that

$$\begin{aligned} |G_\varepsilon(y) - G_\varepsilon(x)| &= L\,|g_k(y) - g_k(x)| \leq L2^N(y-x) \\ &\leq L2^N(y-x)^{1-\alpha}(y-x)^\alpha \leq L2^{N\alpha}(y-x)^\alpha \leq M(y-x)^\alpha. \end{aligned}$$

If $x \in I_k$ and $y \in I_{k'}$ with $k' \geq k+1$,

$$\left(y - 2^{-N}k'\right) + \left(2^{-N}(k+1) - x\right) \leq 2^{-N+1} \wedge (y-x)$$

and since g vanishes at 0 and 1,

$$\begin{aligned} |G_\varepsilon(y) - G_\varepsilon(x)| &= L\,|\varepsilon_{k'+1} g_{k'}(y) - \varepsilon_{k+1} g_k(x)| \leq L\,|g_{k'}(y)| + L\,|g_k(x)| \\ &= L\left|g_{k'}(y) - g_{k'}(2^{-N}k')\right| + L\left|g_k(2^{-N}(k+1)) - g_k(x)\right| \\ &\leq L2^N\left[y - 2^{-N}k' + 2^{-N}(k+1) - x\right]^{1-\alpha+\alpha} \\ &\leq L2^N 2^{(1-\alpha)(-N+1)}(y-x)^\alpha = L2^{(N-1)\alpha+1}(y-x)^\alpha \end{aligned}$$

and the conclusion follows from the fact that $L \leq 2^{-[(N-1)\alpha+1]}M$. □

We use the following version of Assouad's lemma.

Lemma 2.7.1 (Assouad's Lemma)**.** *Let $\mathcal{P}$ be a family of probabilities on a measurable space $(\mathscr{X}, \mathcal{X})$. Assume that for some integer $d \geq 1$, $\mathcal{P}$ contains a subset of the form $\mathcal{C} = \{P_\varepsilon,\ \varepsilon \in \{-1,1\}^d\}$ with the following properties:*

(i) there exists $\eta > 0$ such that for all $\varepsilon, \varepsilon' \in \{-1,1\}^d$

$$h^2\left(P_\varepsilon, P_{\varepsilon'}\right) \geq \eta\delta(\varepsilon,\varepsilon') \quad \text{with} \quad \delta(\varepsilon,\varepsilon') = \sum_{j=1}^{d} \mathbb{1}_{\varepsilon_j \neq \varepsilon'_j}$$

(ii) there exists a constant $a \in [0, 1/2]$ such that

$$h^2\left(P_\varepsilon, P_{\varepsilon'}\right) \leq \frac{a}{n} \quad \text{for all } \varepsilon, \varepsilon' \in \{-1,1\}^d \text{ satisfying } \delta(\varepsilon,\varepsilon') = 1.$$

Then for all measurable mappings $\widehat{P} : \mathscr{X}^n \to \mathcal{P}$,

$$\sup_{P\in\mathcal{P}} \mathbb{E}_{\mathbf{P}}\left[h^2(P, \widehat{P}(\boldsymbol{X}))\right] \geq \frac{d\eta}{8} \max\left\{1 - \sqrt{2a}, (1-a/n)^{2n}\right\}, \tag{2.7.2}$$

where $\mathbb{E}_{\mathbf{P}}$ denotes the expectation with respect to a random variable $\boldsymbol{X} = (X_1, \ldots, X_n)$ with distribution $\mathbf{P} = P^{\otimes n}$.

Proof. Given a probability P on $(\mathscr{X}, \mathcal{X})$, let $\overline{\varepsilon}$ be a minimizer over $\{-1,1\}^d$ of the mapping $\varepsilon \mapsto h^2(P, P_\varepsilon)$. By definition of $\overline{\varepsilon}$, for all $\varepsilon \in \{-1,1\}^d$

$$h^2(P_\varepsilon, P_{\overline{\varepsilon}}) \leq 2\left(h^2(P, P_\varepsilon) + h^2(P, P_{\overline{\varepsilon}})\right) \leq 4h^2(P, P_\varepsilon).$$

Hence by (i), for all $\varepsilon \in \{-1,1\}^d$

$$h^2(P_\varepsilon, P) \geq \frac{\eta}{4}\delta(\varepsilon,\overline{\varepsilon}) = \sum_{i=1}^{d}\left[\frac{1+\varepsilon_i}{2}\ell_i(P) + \frac{1-\varepsilon_i}{2}\ell'_i(P)\right]$$

with $\ell_i(P) = (\eta/4)\mathbb{1}_{\overline{\varepsilon}_i=-1}$ and $\ell'_i(P) = (\eta/4)\mathbb{1}_{\overline{\varepsilon}_i=+1}$ for $i \in \{1,\ldots,d\}$. The result follows by applying the version of Assouad's lemma that can be found in Birgé (1986) with $\beta_i = a/n$ for all $i \in \{1,\ldots,d\}$, $\alpha = \eta/4$ and the change of notation from $\varepsilon \in \{-1,1\}$ to $\varepsilon \in \{0,1\}$. □

2.7.3 Proof of Proposition 2.4.1

Since the statistical model $\overline{\mathscr{Q}} = \{R_\gamma = Q_{v^{-1}(\gamma)},\ \gamma \in J\}$ is regular with constant Fisher information equal to 8, by applying Theorem 7.6 in Ibragimov and Has'minskiĭ (1981)[page 81] we obtain that

$$h^2\left(R_\gamma, R_{\gamma'}\right) \leq \left(\gamma' - \gamma\right)^2 \quad \text{for all } \gamma, \gamma' \in J$$

and for any compact subset K of J, there exists a constant $c_K > 0$

$$h^2\left(R_\gamma, R_{\gamma'}\right) \geq c_K^2\left(\gamma' - \gamma\right)^2 \quad \text{for all } \gamma, \gamma' \in K.$$

The result follows by substituting $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$ to γ and γ' respectively and then integrating with respect to P_W.

2.7.4 Proof of Proposition 2.4.2

For $\boldsymbol{\gamma} \in \mathcal{H}_\alpha(M)$ and $j \in \{1, \ldots, D\}$, let $\gamma_j = D\int_{I_j} \boldsymbol{\gamma}(w)dw$ and $\overline{\boldsymbol{\gamma}} = \sum_{j=1}^D \gamma_j \mathbb{1}_{I_j}$. Since $\boldsymbol{\gamma}$ takes its values in J, $\gamma_j \in J$ for all $j \in \{1, \ldots, D\}$ and $\overline{\boldsymbol{\gamma}} = \sum_{j=1}^D \gamma_j \mathbb{1}_{I_j} \in \overline{\boldsymbol{S}}$. Since for all $w \in I_j$, $|\boldsymbol{\gamma}(w) - \overline{\boldsymbol{\gamma}}(w)| \leq \sup_{|w-w'|\leq 1/D} |\boldsymbol{\gamma}(w) - \boldsymbol{\gamma}(w')| \leq MD^{-\alpha}$ and $\boldsymbol{S}$ is dense in $\overline{\boldsymbol{S}}$ with respect to the supremum norm

$$\begin{aligned}\sup_{\boldsymbol{\gamma}\in\mathcal{H}_\alpha(M)} \inf_{\overline{\boldsymbol{\gamma}}\in\boldsymbol{S}} \|\boldsymbol{\gamma} - \overline{\boldsymbol{\gamma}}\|_2 &\leq \sup_{\boldsymbol{\gamma}\in\mathcal{H}_\alpha(M)} \inf_{\overline{\boldsymbol{\gamma}}\in\boldsymbol{S}} \|\boldsymbol{\gamma} - \overline{\boldsymbol{\gamma}}\|_\infty \\ &= \sup_{\boldsymbol{\gamma}\in\mathcal{H}_\alpha(M)} \inf_{\overline{\boldsymbol{\gamma}}\in\overline{\boldsymbol{S}}} \|\boldsymbol{\gamma} - \overline{\boldsymbol{\gamma}}\|_\infty \leq MD^{-\alpha}.\end{aligned}$$

Using (2.4.4) and the fact that the data $X_1, \ldots, X_n$ are i.i.d., we deduce that for all functions $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$ with values in J,

$$\mathbf{h}^2(\mathbf{R}_{\boldsymbol{\gamma}}, \mathbf{R}_{\boldsymbol{\gamma}'}) = nh^2(R_{\boldsymbol{\gamma}}, R_{\boldsymbol{\gamma}'}) \leq n\kappa^2 \left\|\boldsymbol{\gamma} - \boldsymbol{\gamma}'\right\|_2^2 \leq n\kappa^2 \left\|\boldsymbol{\gamma} - \boldsymbol{\gamma}'\right\|_\infty^2$$

and by applying Corollary 2.3.1 with $V = D + 1$ we derive that

$$\begin{aligned}&\sup_{\boldsymbol{\gamma}^\star\in\mathcal{H}_\alpha(M)} \mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}})\right] \\ &\qquad \leq C'\left[\sup_{\boldsymbol{\gamma}^\star\in\mathcal{H}_\alpha(M)} \inf_{\overline{\boldsymbol{\gamma}}\in\boldsymbol{S}} h^2(R_{\boldsymbol{\gamma}^\star}, R_{\overline{\boldsymbol{\gamma}}}) + \frac{V}{n}\left[1 + \log_+(n/V)\right]\right] \\ &\qquad \leq C'\left[\kappa^2 \sup_{\boldsymbol{\gamma}^\star\in\mathcal{H}_\alpha(M)} \inf_{\overline{\boldsymbol{\gamma}}\in\boldsymbol{S}} \|\boldsymbol{\gamma}^\star - \overline{\boldsymbol{\gamma}}\|_2^2 + \frac{V}{n}\left[1 + \log_+(n/V)\right]\right] \\ &\qquad \leq C'\left[\kappa^2 M^2 D^{-2\alpha} + \frac{D+1}{n}\log(en)\right].\end{aligned}$$

Let us set $L_n = \log(en)$. With our choice of $D \geq 1$,

$$D - 1 < \left(\frac{\kappa^2 M^2 n}{L_n}\right)^{\frac{1}{1+2\alpha}} \leq D$$

hence $\kappa^2 M^2 D^{-2\alpha} \leq DL_n/n$, $D < 1 + (\kappa^2 M^2 n/L_n)^{\frac{1}{1+2\alpha}}$ and the result follows from the inequalities

$$\kappa^2 M^2 D^{-2\alpha} + \frac{(D+1)L_n}{n} \leq 2\frac{DL_n}{n} + \frac{L_n}{n} \leq 2\left[\frac{(\kappa M)^{1/\alpha} L_n}{n}\right]^{\frac{2\alpha}{1+2\alpha}} + \frac{3L_n}{n}.$$

2.7.5 Proof of Proposition 2.4.3

Let a_0 be the middle of the interval K of length $2\overline{L}$. Given $N \geq 1$, $L > 0$ and $\varepsilon \in \{-1,1\}^{2^N}$, we define $\boldsymbol{\gamma}_\varepsilon = a_0 + G_\varepsilon$ with G_ε defined by (2.7.1). Provided that $L \leq \overline{L} \wedge L_0$ with $L_0 = 2^{-[(N-1)\alpha+1]}M$, the functions $\boldsymbol{\gamma}_\varepsilon$ takes their values in $K \subset J$ and satisfies (2.4.1) and consequently belongs to $\mathcal{H}_\alpha(M)$ for all $\varepsilon \in \{-1,1\}^{2^N}$. Set $R_\varepsilon = R_{\boldsymbol{\gamma}_\varepsilon}$ for all $\varepsilon \in \{-1,1\}^{2^N}$. Let us denote by $P_{\boldsymbol{\gamma}} = R_{\boldsymbol{\gamma}} \cdot P_W$ the probability associated to $R_{\boldsymbol{\gamma}}$ and write P_ε for $P_{\boldsymbol{\gamma}_\varepsilon}$ for short. Integrating the inequalities (2.4.4) and (2.4.5) with respect to P_W and using that for all $\varepsilon, \varepsilon' \in \{-1,1\}^{2^N}$, $\|G_\varepsilon - G_{\varepsilon'}\|_2 = \|\boldsymbol{\gamma}_\varepsilon - \boldsymbol{\gamma}_{\varepsilon'}\|_2$ we obtain that

$$c_K^2 \|G_\varepsilon - G_{\varepsilon'}\|_2^2 \leq h^2(R_\varepsilon, R_{\varepsilon'}) \leq \kappa^2 \|G_\varepsilon - G_{\varepsilon'}\|_2^2 .$$

Let us set $\Lambda = \{0, \ldots, 2^N - 1\}$. Since P_W is the uniform distribution and the supports of the functions $g_k : x \mapsto g(2^N x - k)$ for $k \in \Lambda$ are disjoint, we obtain that for all $\varepsilon, \varepsilon' \in \{-1,1\}^{2^N}$

$$\begin{aligned}\|G_\varepsilon - G_{\varepsilon'}\|_2^2 &= L^2 \sum_{k\in\Lambda} \int_{I_k} \left(\varepsilon_{k+1} g_k(x) - \varepsilon'_{k+1} g_k(x)\right)^2 dx \\ &= L^2 \sum_{k\in\Lambda} \left|\varepsilon_{k+1} - \varepsilon'_{k+1}\right|^2 \int_{I_k} g_k^2(x)dx = 4L^2 2^{-N} \|g\|_2^2 \, \delta(\varepsilon, \varepsilon')\end{aligned}$$

and consequently, provided that L satisfies

$$L \leq \overline{L} \wedge L_0 \wedge \left(4\kappa \|g\|_2 \sqrt{2^{-(N-1)}n}\right)^{-1} \tag{2.7.3}$$

the family of probabilities $\mathscr{C} = \{P_\varepsilon,\ \varepsilon \in \{-1,1\}^{|\Lambda|}\}$ is a subset of $\mathcal{P} = \{P_{\boldsymbol{\gamma}},\ \boldsymbol{\gamma} \in \mathcal{H}_\alpha(M)\}$ that fulfils the assumptions of Lemma 2.7.1 with $d = 2^N$,

$$\eta = 4c_K^2 L^2 2^{-N} \|g\|_2^2 \quad \text{and} \quad a = 4n\kappa^2 L^2 2^{-N} \|g\|_2^2 \leq 1/8.$$

We derive from (2.7.2) that

$$\mathcal{R}_n(\mathcal{H}_\alpha(M)) \geq \frac{c_K^2 \|g\|_2^2 L^2}{2} \left(1 - \sqrt{2a}\right) \geq \frac{c_K^2 \|g\|_2^2 L^2}{4}. \tag{2.7.4}$$

If $\kappa^2 \|g\|_2^2 M^2 n > 1/8$, we choose $N \geq 2$ such that

$$2^N \geq \left(2^{2(2+\alpha)} \kappa^2 \|g\|_2^2 M^2 n\right)^{1/(1+2\alpha)} > 2^{N-1}$$

and $N = 1$ otherwise. In any case, our choice of N satisfies

$$L_0 = 2^{-[(N-1)\alpha+1]} M \leq \left(4\kappa \|g\|_2 \sqrt{2^{-(N-1)}n}\right)^{-1}.$$

When $N \geq 2$,

$$\begin{aligned}L_0^2 = 2^{-2\alpha(N-1)-2} M^2 &\geq \frac{M^2}{4} \left(2^{2(2+\alpha)} \kappa^2 \|g\|_2^2 M^2 n\right)^{-\frac{2\alpha}{1+2\alpha}} \\ &= \left(\frac{M^{1/\alpha}}{2^{2\alpha+6+1/\alpha} \kappa^2 \|g\|_2^2 n}\right)^{\frac{2\alpha}{1+2\alpha}} = L_1^2,\end{aligned}$$

while $L_0 = M/2$ when $N = 1$. The choice $L = \overline{L} \wedge L_1 \wedge (M/2)$ satisfies (2.7.3) and we deduce from the equalities

$$h^2(R_\varepsilon, R_{\varepsilon'}) = \int_{\mathscr{W}} h^2\left(R_{\boldsymbol{\gamma}_\varepsilon(w)}, R_{\boldsymbol{\gamma}_{\varepsilon'}(w)}\right) dP_W(w) = h^2(P_\varepsilon, P_{\varepsilon'})$$

and (2.7.4) that

$$\mathcal{R}_n(\mathcal{H}_\alpha(M)) \geq \frac{c_K^2 \|g\|_2^2}{4} \left[\left(\frac{M^{1/\alpha}}{2^{2\alpha+6+1/\alpha} \kappa^2 \|g\|_2^2 n} \right)^{\frac{2\alpha}{1+2\alpha}} \bigwedge \left(\frac{M^2}{4} \right) \bigwedge \overline{L}^2 \right].$$

The conclusion follows by choosing $g(x) = x\mathbb{1}_{[0,1/2]} + (1-x)\mathbb{1}_{[1/2,1]}$ which satisfies $\|g\|_2^2 = 1/12$.

2.7.6 Proof of Proposition 2.4.4

When the Poisson family is parametrized by the mean, given two functions $\boldsymbol{\gamma}, \boldsymbol{\gamma}'$ mapping $\mathscr{W} = [0,1]$ into $J = (0, +\infty)$, The Hellinger-type distance $h^2(R_{\boldsymbol{\gamma}}, R_{\boldsymbol{\gamma}'})$ can be written as

$$h^2(R_{\boldsymbol{\gamma}}, R_{\boldsymbol{\gamma}'}) = \int_{\mathscr{W}} \left[1 - e^{-\left(\sqrt{\boldsymbol{\gamma}(w)} - \sqrt{\boldsymbol{\gamma}'(w)}\right)^2/2} \right] dP_W(w). \tag{2.7.5}$$

Using that for all $x \in [0,1]$, $(1-e^{-1})x \leq 1 - e^{-x} \leq x$, we deduce from (2.7.5) that

$$\frac{1}{2}(1-e^{-1}) \left\| \sqrt{\boldsymbol{\gamma}} - \sqrt{\boldsymbol{\gamma}'} \right\|_2^2 \leq h^2(R_{\boldsymbol{\gamma}}, R_{\boldsymbol{\gamma}'}) \leq \frac{1}{2} \left\| \sqrt{\boldsymbol{\gamma}} - \sqrt{\boldsymbol{\gamma}'} \right\|_2^2 \tag{2.7.6}$$

whenever $\left\| \sqrt{\boldsymbol{\gamma}} - \sqrt{\boldsymbol{\gamma}'} \right\|_\infty \leq 1$.

Let N be some positive integer, L some positive number and g a 1-Lipschitz function supported on $[0,1]$ with values in $[-b, b]$. Let us set $\Lambda = \{0, \ldots, 2^N - 1\}$ and for $\varepsilon \in \{-1,1\}^{|\Lambda|}$, G_ε the function defined by (2.7.1) and $\boldsymbol{\gamma}_\varepsilon = L + G_\varepsilon$. Under our assumption on g, $\boldsymbol{\gamma}_\varepsilon$ takes its values in $[(1-b)L, (1+b)L]$ and by Proposition 2.7.3, $\boldsymbol{\gamma}_\varepsilon$ satisfies (2.4.1) provided that $L \leq 2^{-[(N-1)\alpha+1]}M$. Hence, under the conditions $L \leq 2^{-[(N-1)\alpha+1]}M$ and $b < 1$, $\boldsymbol{\gamma}_\varepsilon$ belongs to $\mathcal{H}_\alpha(M)$ for all $\varepsilon \in \{-1,1\}^{|\Lambda|}$. For all $\varepsilon, \varepsilon' \in \{-1,1\}^{|\Lambda|}$,

$$\frac{|G_\varepsilon - G_{\varepsilon'}|}{2\sqrt{(1+b)L}} \leq |\sqrt{\boldsymbol{\gamma}_\varepsilon} - \sqrt{\boldsymbol{\gamma}_{\varepsilon'}}| = \frac{|\boldsymbol{\gamma}_\varepsilon - \boldsymbol{\gamma}_{\varepsilon'}|}{\sqrt{\boldsymbol{\gamma}_\varepsilon} + \sqrt{\boldsymbol{\gamma}_{\varepsilon'}}} \leq \frac{|G_\varepsilon - G_{\varepsilon'}|}{2\sqrt{(1-b)L}},$$

and

$$|\sqrt{\boldsymbol{\gamma}_\varepsilon} - \sqrt{\boldsymbol{\gamma}_{\varepsilon'}}| \leq \sqrt{(1+b)L} - \sqrt{(1-b)L} = \left[\sqrt{1+b} - \sqrt{1-b}\right] \sqrt{L}.$$

In particular, $\left\| \sqrt{\boldsymbol{\gamma}_\varepsilon} - \sqrt{\boldsymbol{\gamma}_{\varepsilon'}} \right\|_\infty \leq 1$ for

$$L \leq \left(\sqrt{1+b} - \sqrt{1-b}\right)^{-2} = \frac{1+\sqrt{1-b^2}}{2b^2} = L_0$$

and, writing R_ε for $R_{\boldsymbol{\gamma}_\varepsilon}$ for short, it follows from (2.7.6) that

$$\frac{(1-e^{-1})}{8(1+b)L}\,\|G_\varepsilon - G_{\varepsilon'}\|_2^2 \le h^2(R_\varepsilon, R_{\varepsilon'}) \le \frac{1}{8(1-b)L}\,\|G_\varepsilon - G_{\varepsilon'}\|_2^2\,. \tag{2.7.7}$$

Since P_W is the uniform distribution and the supports of the functions $g_k : x \mapsto g(2^N x - k)$ for $k \in \Lambda$ are disjoint, we obtain that for all $\varepsilon, \varepsilon' \in \{-1,1\}^{|\Lambda|}$

$$\begin{aligned}\|G_\varepsilon - G_{\varepsilon'}\|_2^2 &= L^2 \sum_{k\in\Lambda} \int_{I_k} \left(\varepsilon_{k+1} g_k(x) - \varepsilon'_{k+1} g_k(x)\right)^2 dx \\ &= L^2 \sum_{k\in\Lambda} \left|\varepsilon_{k+1} - \varepsilon'_{k+1}\right|^2 \int_{I_k} g_k^2(x)dx = 4L^2 2^{-N}\,\|g\|_2^2\,\delta(\varepsilon,\varepsilon').\end{aligned}$$

Let us denote by $P_{\boldsymbol{\gamma}} = R_{\boldsymbol{\gamma}} \cdot P_W$ the probability associated to $R_{\boldsymbol{\gamma}}$ and write P_ε for $P_{\boldsymbol{\gamma}_\varepsilon}$ for short. We deduce from (2.7.7) that provided that L and b satisfy

$$L \le \left(2^{-[(N-1)\alpha+1]} M\right) \bigwedge \frac{1+\sqrt{1-b^2}}{2b^2} \bigwedge \frac{(1-b)2^{N-3}}{\|g\|_2^2\, n}, \tag{2.7.8}$$

the family of probabilities $\mathscr{C} = \{P_\varepsilon,\ \varepsilon \in \{-1,1\}^{|\Lambda|}\}$ is a subset of $\{P_{\boldsymbol{\gamma}},\ \boldsymbol{\gamma} \in \mathcal{H}_\alpha(M)\}$ that fulfils the assumptions of Assouad's lemma (Lemma 2.7.1) with $d = |\Lambda| = 2^N$,

$$\eta = \frac{(1-e^{-1})L2^{-(N+1)}\,\|g\|_2^2}{1+b} \quad \text{and} \quad a = \frac{nL2^{-N}\,\|g\|_2^2}{1-b} \in [0, 1/8].$$

We derive from the equalities

$$h^2(R_\varepsilon, R_{\varepsilon'}) = \int_{\mathscr{W}} h^2\left(R_{\boldsymbol{\gamma}_\varepsilon(w)}, R_{\boldsymbol{\gamma}_{\varepsilon'}(w)}\right) dP_W(w) = h^2(P_\varepsilon, P_{\varepsilon'})$$

and (2.7.2) that

$$\mathcal{R}_n(\mathcal{H}_\alpha(M)) \ge \frac{(1-e^{-1})\,\|g\|_2^2\, L}{16(1+b)}\left(1-\sqrt{2a}\right) \ge \frac{(1-e^{-1})\,\|g\|_2^2\, L}{32(1+b)}. \tag{2.7.9}$$

If $\|g\|_2^2\, Mn > (1-b)/2$, we choose $N \ge 2$ such that

$$2^N \ge \left[\frac{2^{2+\alpha}\,\|g\|_2^2\, Mn}{1-b}\right]^{\frac{1}{1+\alpha}} > 2^{N-1}.$$

Otherwise, we choose $N = 1$. Note that in any case,

$$2^{-[(N-1)\alpha+1]} M \le \frac{(1-b)2^{N-3}}{n\,\|g\|_2^2}.$$

Besides, if $N \ge 2$

$$\begin{aligned}2^{-[(N-1)\alpha+1]} M = 2^{-1} M 2^{-(N-1)\alpha} &\ge 2^{-1} M \left[\frac{2^{2+\alpha}\,\|g\|_2^2\, Mn}{1-b}\right]^{-\frac{\alpha}{1+\alpha}} \\ &= \left(\frac{(1-b)M^{\frac{1}{\alpha}}}{2^{3+\alpha+1/\alpha}\,\|g\|_2^2\, n}\right)^{\frac{\alpha}{1+\alpha}} = L_1\end{aligned}$$

while for $2^{-[(N-1)\alpha+1]}M = M/2$ for $N = 1$. Finally, we choose $L = L_0 \wedge L_1 \wedge (M/2)$, which satisfies (2.7.8), and we derive from (2.7.9) that

$$\mathcal{R}_n(\mathcal{H}_\alpha(M)) \geq \frac{(1-e^{-1})\|g\|_2^2}{32(1+b)} \left[\left(\frac{(1-b)M^{\frac{1}{\alpha}}}{2^{3+\alpha+1/\alpha}\,\|g\|_2^2\, n} \right)^{\frac{\alpha}{1+\alpha}} \wedge \frac{M}{2} \wedge \frac{1+\sqrt{1-b^2}}{b^2} \right].$$

The conclusion follows by taking $g(x) = x\mathbb{1}_{[0,1/2]} + (1-x)\mathbb{1}_{[1/2,1]}$ for which $b = 1/2$ and $\|g\|_2^2 = 1/12$.

2.7.7 Proof of Proposition 2.4.5

Let $\alpha \in (0,1]$ and $M > 0$. For a function $\boldsymbol{\gamma} \in \mathcal{H}_\alpha(M)$ taking values in $J = (0, +\infty)$, we set $\gamma_j = D\int_{I_j} \boldsymbol{\gamma}(w)dw$, for $j \in \{1, \dots, D\}$ and $\overline{\boldsymbol{\gamma}} = \sum_{j=1}^D \gamma_j \mathbb{1}_{I_j}$. As an immediate consequence, $\gamma_j \in J$ for all $j \in \{1, \dots, D\}$ and $\overline{\boldsymbol{\gamma}} \in \overline{\boldsymbol{S}}$. Since for all $w \in I_j$, with the fact that $\boldsymbol{\gamma} \in \mathcal{H}_\alpha(M)$

$$|\boldsymbol{\gamma}(w) - \overline{\boldsymbol{\gamma}}(w)| \leq \sup_{|w-w'|\leq 1/D} \left|\boldsymbol{\gamma}(w) - \boldsymbol{\gamma}(w')\right| \leq MD^{-\alpha}$$

and $\boldsymbol{S}$ is dense in $\overline{\boldsymbol{S}}$ with respect to the supremum norm, we derive

$$\begin{aligned}
\sup_{\boldsymbol{\gamma}\in\mathcal{H}_\alpha(M)} \inf_{\overline{\boldsymbol{\gamma}}\in\boldsymbol{S}} \left\|\sqrt{\boldsymbol{\gamma}} - \sqrt{\overline{\boldsymbol{\gamma}}}\right\|_2 &\leq \sup_{\boldsymbol{\gamma}\in\mathcal{H}_\alpha(M)} \inf_{\overline{\boldsymbol{\gamma}}\in\boldsymbol{S}} \left\|\sqrt{\boldsymbol{\gamma}} - \sqrt{\overline{\boldsymbol{\gamma}}}\right\|_\infty \\
&\leq \sup_{\boldsymbol{\gamma}\in\mathcal{H}_\alpha(M)} \inf_{\overline{\boldsymbol{\gamma}}\in\boldsymbol{S}} \sqrt{\|\boldsymbol{\gamma} - \overline{\boldsymbol{\gamma}}\|_\infty} \\
&= \sup_{\boldsymbol{\gamma}\in\mathcal{H}_\alpha(M)} \inf_{\overline{\boldsymbol{\gamma}}\in\overline{\boldsymbol{S}}} \sqrt{\|\boldsymbol{\gamma} - \overline{\boldsymbol{\gamma}}\|_\infty} \\
&\leq \sqrt{M}D^{-\frac{\alpha}{2}}.
\end{aligned}$$

Using the fact that the data $X_1, \dots, X_n$ are i.i.d. and $1 - e^{-x} \leq x$ for all $x \in [0, +\infty)$, we deduce that for all functions $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$ with values in $J = (0, +\infty)$,

$$\begin{aligned}
\mathbf{h}^2(\mathbf{R}_{\boldsymbol{\gamma}}, \mathbf{R}_{\boldsymbol{\gamma}'}) = nh^2(R_{\boldsymbol{\gamma}}, R_{\boldsymbol{\gamma}'}) &= n\int_{\mathscr{W}} \left[1 - e^{-\left(\sqrt{\boldsymbol{\gamma}(w)} - \sqrt{\boldsymbol{\gamma}'(w)}\right)^2/2}\right] dP_W(w) \\
&\leq \frac{n}{2}\left\|\sqrt{\boldsymbol{\gamma}} - \sqrt{\boldsymbol{\gamma}'}\right\|_2^2. \qquad (2.7.10)
\end{aligned}$$

Applying Corollary 2.3.1 with $V = D + 1$ together with (2.7.10), we obtain that

$$\begin{aligned}
\sup_{\boldsymbol{\gamma}^\star\in\mathcal{H}_\alpha(M)} \mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}})\right] &\leq C' \left[\sup_{\boldsymbol{\gamma}^\star\in\mathcal{H}_\alpha(M)} \inf_{\overline{\boldsymbol{\gamma}}\in\boldsymbol{S}} h^2(R_{\boldsymbol{\gamma}^\star}, R_{\overline{\boldsymbol{\gamma}}}) + \frac{V}{n}\left[1 + \log_+(n/V)\right]\right] \\
&\leq C' \left[\frac{1}{2}\sup_{\boldsymbol{\gamma}^\star\in\mathcal{H}_\alpha(M)} \inf_{\overline{\boldsymbol{\gamma}}\in\boldsymbol{S}} \left\|\sqrt{\boldsymbol{\gamma}^\star} - \sqrt{\overline{\boldsymbol{\gamma}}}\right\|_2^2 + \frac{V}{n}\left[1 + \log_+(n/V)\right]\right] \\
&\leq C' \left[\frac{1}{2}MD^{-\alpha} + \frac{D+1}{n}\log(en)\right].
\end{aligned}$$

Let us set $L_n = \log(en)$. With our choice of $D \geq 1$,

$$D - 1 < \left(\frac{Mn}{2L_n}\right)^{\frac{1}{1+\alpha}} \leq D$$

hence $MD^{-\alpha}/2 \leq DL_n/n$, $D < 1 + [Mn/(2L_n)]^{1/(1+\alpha)}$ and the result follows from the inequalities

$$\frac{MD^{-\alpha}}{2} + \frac{(D+1)L_n}{n} \leq 2\frac{DL_n}{n} + \frac{L_n}{n} \leq 2\left[\frac{(M/2)^{1/\alpha}L_n}{n}\right]^{\frac{\alpha}{1+\alpha}} + \frac{3L_n}{n}.$$

Chapter 3

Estimation by model selection

3.1 Introduction

We observe n independent pairs of random variables $X_1 = (W_1, Y_1), \ldots, X_n = (W_n, Y_n)$ with values in a measurable product space $(\mathscr{X}, \mathcal{X}) = (\mathscr{W} \times \mathscr{Y}, \mathcal{W} \otimes \mathcal{Y})$. Recall that, as we have introduced in Section 2.2, $\mathscr{T}$ denotes the set of all probabilities on $(\mathscr{Y}, \mathcal{Y})$ which we equip with the Borel σ-algebra $\mathcal{T}$ associated to the total variation distance and the notation $\mathscr{Q}_{\mathscr{W}}$ stands for the set of all measurable mappings (conditional probabilities) from $(\mathscr{W}, \mathcal{W})$ into $(\mathscr{T}, \mathcal{T})$. We assume that for each $i \in \{1, \ldots, n\}$, the conditional distribution $Q_i^\star(w_i)$ of Y_i given $W_i = w_i$ exists and is given by the value at w_i of some measurable function $Q_i^\star \in \mathscr{Q}_{\mathscr{W}}$. Our goal is to estimate the n-tuple $\mathbf{Q}^\star = (Q_1^\star, \ldots, Q_n^\star)$. We do as if there exists an unknown function $\boldsymbol{\gamma}^\star$ on $\mathscr{W}$ such that for each $i \in \{1, \ldots, n\}$, the conditional distribution of Y_i given $W_i = w_i$ belongs to a one-parameter exponential family with parameter $\boldsymbol{\gamma}^\star(w_i) \in \mathbb{R}$. When such a $\boldsymbol{\gamma}^\star$ does exist, the above statistical setting includes binary, Gaussian and Poisson regressions and exponential multiplicative regression, among many others.

In Chapter 2, we have proposed a robust procedure based on the ρ-estimation to estimate $\mathbf{Q}^\star$. The approach is restricted to the case of a single model. Up to a numerical constant, the risk of our ρ-estimator $\widehat{\boldsymbol{\gamma}}$ within the constructed 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 the potential function $\boldsymbol{\gamma}^\star$.

In this chapter, we consider the same estimation problem, i.e. estimating the conditional distributions $Q_i^\star(w_i)$ of Y_i given $W_i = w_i$, by model selection. For an exponential family $\overline{\mathscr{Q}}$, we focus on its general form of parametrization, i.e. $\overline{\mathscr{Q}} = \{R_\gamma = \overline{r}_\gamma \cdot \mu,\ \gamma \in J\}$

where the densities (with respect to μ) are of the form, for all $y \in \mathscr{Y}$ and $\gamma \in J$

$$\overline{r}_\gamma(y) = e^{u(\gamma)S(y)-B(\gamma)}a(y) \text{ where } B(\gamma) = \log\left[\int_{\mathscr{Y}} e^{u(\gamma)S(y)}a(y)d\mu(y)\right], \tag{3.1.1}$$

S 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 J and a is a nonnegative function on $\mathscr{Y}$. For convenience, we denote

$$r_\gamma(y) = e^{u(\gamma)S(y)-B(\gamma)}, \quad \text{for all } y \in \mathscr{Y} \text{ and } \gamma \in J, \tag{3.1.2}$$

and rewrite $\overline{\mathscr{Q}} = \{R_\gamma = r_\gamma \cdot \nu,\ \gamma \in J\}$ which is exact the form described in (2.3.12).

We propose a model selection procedure based on ρ-estimation to estimate the conditional distributions $\mathbf{Q}^\star = (Q_1^\star, \ldots, Q_n^\star)$ and establish non-asymptotic exponential inequalities for the upper deviations between the resulted estimator and the truth $\mathbf{Q}^\star$. Our approach is still based on the presumption that there exists an unknown $\boldsymbol{\gamma}^\star$ on $\mathscr{W}$ belonging to some of our models such that $Q_i^\star(w_i)$ is of the form $R_{\boldsymbol{\gamma}^\star(w_i)}$ for all $i \in \{1, \ldots, n\}$. However, our approach is not restricted to this assumption as we have emphasised in former chapters. Our estimator takes the form of a mapping $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}} : \mathbf{w} = (w_1, \ldots, w_n) \in \mathscr{W}^n \mapsto (R_{\widehat{\boldsymbol{\gamma}}(w_1)}, \ldots, R_{\widehat{\boldsymbol{\gamma}}(w_n)})$ with values in $\overline{\mathscr{Q}}^n$, where $\widehat{\boldsymbol{\gamma}}$ is a (random) function from $\mathscr{W}$ into J. In particular, when $\mathbf{Q}^\star = \mathbf{R}_{\boldsymbol{\gamma}^\star}$ for some (deterministic) function $\boldsymbol{\gamma}^\star : \mathscr{W} \to I$, $\widehat{\boldsymbol{\gamma}}$ provides an estimator of the so called *regression function* $\boldsymbol{\gamma}^\star$. We also keep to endow $\boldsymbol{\mathscr{Q}}_{\mathscr{W}} = \mathscr{Q}_{\mathscr{W}}^n$ with the Hellinger-type (pseudo) distance $\mathbf{h}$ introduced in (2.2.2). When W_i are i.i.d. with the common distribution P_W and $Q_i^\star = Q^\star$ for all $i \in \{1, \ldots, n\}$, we slightly abuse the notation $h^2(Q^\star, R_{\widehat{\boldsymbol{\gamma}}})$ to measure the distance between $Q^\star$ and $R_{\widehat{\boldsymbol{\gamma}}}$ which is defined as

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

Besides the model selection procedure, we put more attention to the situation when $X_1, \ldots, X_n$ are i.i.d., where several interesting applications arise including adaptation and variable selection problems in exponential families. Also 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. In particular, 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. The above mentioned results are replied on constructing suitable 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 VC dimension bounds can be of independent interest.

The remainder of this chapter is organized as follows. We introduce the estimation procedure in Section 3.2 together with its theoretical properties. We then discuss the adaptive estimation problem in exponential families when the regression function belongs to anisotropic Besov spaces as an application in Section 3.3. 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 3.4, 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 3.5 and variable selection problem in generalized linear models is considered in Section 3.6. Finally, all the proofs of this chapter can be found in Section 3.7.

We end this section by introducing some notations for later use in this chapter. 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 $\mathbf{Q} \in \boldsymbol{\mathcal{Q}}_{\mathscr{W}}$ and a set $\mathbf{A} \subset \boldsymbol{\mathcal{Q}}_{\mathscr{W}}$, we define $\mathbf{h}^2(\mathbf{Q}, \mathbf{A}) = \inf_{\mathbf{Q}' \in \mathbf{A}} \mathbf{h}^2(\mathbf{Q}, \mathbf{Q}')$. 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 this chapter, C denotes positive numerical constant which may vary from line to line.

3.2 An estimation strategy based on model selection

Our model selection approach is based on ρ-estimation. We refer to Baraud and Birgé (2018) and Baraud et al. (2017) for a thorough study of this methodology or Chapter 1 for a brief introduction.

3.2.1 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 J, 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 3.2.1. 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 definitions and more properties of the VC-subgraph class of functions, we refer to Section 1.5. As we have commented in Section 1.5, one property 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 $\overline{\mathbf{\Gamma}}_a = \{\boldsymbol{\gamma} \vee a,\ \boldsymbol{\gamma} \in \overline{\mathbf{\Gamma}}\}$ and $\overline{\mathbf{\Gamma}}^b = \{\boldsymbol{\gamma} \wedge b,\ \boldsymbol{\gamma} \in \overline{\mathbf{\Gamma}}\}$ are also VC-subgraphs on $\mathscr{W}$ with dimension not larger than V. We shall repeatedly use the conclusion through this chapter.

3.2.2 Model selection procedure

We consider $\{\overline{\mathbf{\Gamma}}_m,\ m \in \mathcal{M}\}$ an at most countable family of models satisfying Assumption 3.2.1. 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 $[0, +\infty]$ given by (2.3.1). For any $\boldsymbol{\gamma}, \boldsymbol{\gamma}' \in \mathbf{\Gamma}$, we introduce the function

$$\mathbf{T}(\boldsymbol{X}, \boldsymbol{\gamma}, \boldsymbol{\gamma}') = \sum_{i=1}^{n} \psi\left(\sqrt{\frac{r_{\boldsymbol{\gamma}'(W_i)}(Y_i)}{r_{\boldsymbol{\gamma}(W_i)}(Y_i)}}\right) \tag{3.2.1}$$

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

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

$$\Sigma = \sum_{m \in \mathcal{M}} e^{-\Delta(m)} < +\infty. \tag{3.2.2}$$

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

$$\mathbf{pen}(\boldsymbol{\gamma}) = 10^2 \inf_{\{m\in\mathcal{M}|\boldsymbol{\gamma}\in\mathbf{\Gamma}_m\}} \left[D_n(m) + 4.7\Delta(m)\right], \quad \text{for all } \boldsymbol{\gamma}\in\mathbf{\Gamma}. \tag{3.2.3}$$

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

$$\boldsymbol{\upsilon}(\boldsymbol{X},\boldsymbol{\gamma}) = \sup_{\boldsymbol{\gamma}'\in\mathbf{\Gamma}} \left[\mathbf{T}(\boldsymbol{X},\boldsymbol{\gamma},\boldsymbol{\gamma}') - \mathbf{pen}(\boldsymbol{\gamma}')\right] + \mathbf{pen}(\boldsymbol{\gamma}). \tag{3.2.4}$$

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

$$\mathscr{E}(\boldsymbol{X}) = \left\{\boldsymbol{\gamma}\in\mathbf{\Gamma} \text{ such that } \boldsymbol{\upsilon}(\boldsymbol{X},\boldsymbol{\gamma}) \le \inf_{\boldsymbol{\gamma}'\in\mathbf{\Gamma}} \boldsymbol{\upsilon}(\boldsymbol{X},\boldsymbol{\gamma}') + \frac{\kappa_\rho}{25}\right\}, \tag{3.2.5}$$

where $\kappa_\rho = 280\sqrt{2} + 74$. Finally, the random variable $\widehat{\boldsymbol{\gamma}}(\boldsymbol{X})$ is our estimator of the regression function $\boldsymbol{\gamma}^\star$ and $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}} = (R_{\widehat{\boldsymbol{\gamma}}},\dots,R_{\widehat{\boldsymbol{\gamma}}})$ is our estimator of $\mathbf{Q}^\star$.

As one can observe from the construction procedure, our estimator depends on the choice of the exponential family $\overline{\mathscr{Q}}$, 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 about the distributions of covariates W_i which, therefore, could be unknown. This is one of the feature distinguishing our procedure with the existing ones Antoniadis and Sapatinas (2001), Antoniadis et al. (2001), Sardy et al. (2004) and Brown et al. (2010) in the literature.

3.2.3 The performance of the estimator

Theorem 3.2.1. *Let $\mathscr{Q}_m = \{\mathbf{R}_{\boldsymbol{\gamma}},\ \boldsymbol{\gamma}\in\mathbf{\Gamma}_m\}$ and $\Xi(m) = D_n(m)/4.7 + \Delta(m)$, for all $m\in\mathcal{M}$. Under Assumption 3.2.1, whatever the conditional probabilities $\mathbf{Q}^\star = (Q_1^\star,\dots,Q_n^\star)$ 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 3.2.2 satisfies for any $\xi>0$, with a probability at least $1-\Sigma e^{-\xi}$*

$$\mathbf{h}^2(\mathbf{Q}^\star,\mathbf{R}_{\widehat{\boldsymbol{\gamma}}}) \le \inf_{m\in\mathcal{M}} \left[c_1\mathbf{h}^2(\mathbf{Q}^\star,\mathscr{Q}_m) + c_2\left(\Xi(m) + 1.49 + \xi\right)\right], \tag{3.2.6}$$

where $c_1 = 150$ and $c_2 = 5014$.

The proof of Theorem 3.2.1 is postponed to Section 3.7.1. We shall use (3.2.6) in the forthcoming sections to solve many model selection problems simultaneously. We give some comments on this 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 $\overline{\mathscr{Q}}_m = \{\mathbf{R}_{\boldsymbol{\gamma}},\ \boldsymbol{\gamma}\in\overline{\mathbf{\Gamma}}_m\}$. If for all $m\in\mathcal{M}$, $\mathscr{Q}_m$ is dense in $\overline{\mathscr{Q}}_m$ with respect to the pseudo Hellinger distance $\mathbf{h}$, i.e. $\mathbf{h}(\mathbf{Q}^\star,\overline{\mathscr{Q}}_m) = \mathbf{h}(\mathbf{Q}^\star,\mathscr{Q}_m)$, (3.2.6) is equivalent to

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

where we involve the models $\overline{\mathbf{\Gamma}}_m$ into the deviation bound of our estimator but not its countable subset $\mathbf{\Gamma}_m$ as we derived in (3.2.6). 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 (3.2.6) with respect to ξ leads to

$$\mathbb{E}\left[\mathbf{h}^2(\mathbf{Q}^\star, \mathbf{R}_{\widehat{\boldsymbol{\gamma}}})\right] \leq \inf_{m\in\mathcal{M}} \left[c_1 \mathbf{h}^2(\mathbf{Q}^\star, \boldsymbol{\mathscr{Q}}_m) + c_2\left(\Xi(m) + \Sigma + 1.49\right)\right]. \tag{3.2.7}$$

We note from (3.2.7) 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 $\boldsymbol{\mathscr{Q}}_m$ to $\mathbf{Q}^\star$, 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 $\mathbf{h}(\mathbf{Q}^\star, \boldsymbol{\mathscr{Q}}_m)$ in (3.2.7) 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. $\mathbf{Q}^\star = \mathbf{R}_{\boldsymbol{\gamma}^\star}$ 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 $V_m\left[1 + \log_+\left(n/V_m\right)\right]$. If it is not the case, the risk involves an additional bias term $\mathbf{h}^2(\mathbf{Q}^\star, \boldsymbol{\mathscr{Q}}_m)$ due to a potential model misspecification or data contamination. However, as long as this bias term remains small compared to $V_m\left[1 + \log_+\left(n/V_m\right)\right]$, 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 $Q_i^\star = Q^\star$ for all $i \in \{1, \ldots, n\}$, we deduce from (3.2.7) that for any $Q^\star$ and P_W, our estimator $R_{\widehat{\boldsymbol{\gamma}}}$ satisfies

$$\mathbb{E}\left[h^2(Q^\star, R_{\widehat{\boldsymbol{\gamma}}})\right] \leq c_2\left(c_3 + \Sigma\right) \inf_{m\in\mathcal{M}} \left[h^2(Q^\star, \mathscr{Q}_m) + \frac{\Delta(m)}{n} + \frac{V_m}{n} L_n(m)\right], \tag{3.2.8}$$

where $c_3 = 1940$, $\mathscr{Q}_m = \{R_{\boldsymbol{\gamma}},\ \boldsymbol{\gamma} \in \mathbf{\Gamma}_m\}$ and $L_n(m) = 1 + \log_+\left(n/V_m\right)$.

3.3 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 $Q_i^\star = Q^\star$ for all $i \in \{1, \ldots, n\}$ and consider adaptive estimation in exponential families. The problem is stated as follows.

Let $0 < p, q \leq \infty$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) \in (\mathbb{R}_+^*)^d$ 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 $B^{\boldsymbol{\alpha}}_{p,q}([0,1]^d)$ 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 (2019). For readers who concern the definitions, we refer to 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 $f \in B^{\boldsymbol{\alpha}}_{p,q}([0,1]^d)$, if α_j is large, then f is smooth to the j-th direction.

For a given interval $[v_-, v_+] \subset J$ with $v_- < v_+$, the notation $B^{\boldsymbol{\alpha}}_{p,q}(R, v_-, v_+)$ stands for the collection of functions $f \in B^{\boldsymbol{\alpha}}_{p,q}([0,1]^d, R)$ with $f(\boldsymbol{w}) \in [v_-, v_+]$ for all $\boldsymbol{w} \in [0,1]^d$. We assume that the regression function $\boldsymbol{\gamma}^\star \in B^{\boldsymbol{\alpha}}_{p,q}(R, v_-, v_+)$. 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.

3.3.1 Models construction

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

$$I_j(k_j) = \begin{cases} \left[0, 2^{-s_j}\right] & , \quad k_j = 0, \\ (k_j 2^{-s_j}, (k_j + 1)2^{-s_j}] & , \quad k_j = 1, \dots, 2^{s_j} - 1. \end{cases} \tag{3.3.1}$$

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 $M^{\mathcal{B},d}_{\boldsymbol{s}}$ 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 $(\overline{k}_1, \dots, \overline{k}_d) \in \Psi(s_1) \times \cdots \times \Psi(s_d)$, any $\boldsymbol{\gamma} \in \overline{\boldsymbol{S}}^{\mathcal{B},d}_{(\boldsymbol{s},r)}$ is of the form for all $\boldsymbol{w} = (w_1, \dots, w_d) \in \prod_{j=1}^d I_j(\overline{k}_j)$

$$\boldsymbol{\gamma}(\boldsymbol{w}) = \sum_{(r_1,\dots,r_d) \in \{0,\dots,r\}^d} \gamma_{(r_1,\dots,r_d)} \prod_{j=1}^d {w_j}^{r_j}, \tag{3.3.2}$$

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

Recall that in our setting $\boldsymbol{\gamma}^\star$ takes values in some non-trivial interval J which may vary from the choice of the exponential family and the choice of parametrization. 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 $[v_-, v_+] \subset J$. 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 (3.3.2) 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 3.3.1. *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 Proposition 1.5.1, 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 3.2.1 with

$$V_{(\boldsymbol{s},r)} = (r+1)^d \prod_{j=1}^d 2^{s_j} + 1. \tag{3.3.3}$$

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

$$\Delta(\boldsymbol{s},r) = \log(8d) \prod_{j=1}^d 2^{s_j} + r. \tag{3.3.4}$$

We have the following result which shows inequality (3.2.2) is satisfied with the weights defined by (3.3.4).

Lemma 3.3.2. *For each $(\boldsymbol{s},r) \in \mathcal{M}$, let the weight be assigned by (3.3.4). Then*

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

We denote $M^{\mathcal{B},d} = \cup_{\boldsymbol{s}\in\mathbb{N}^d} M^{\mathcal{B},d}_{\boldsymbol{s}}$. 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 $\overline{\boldsymbol{\Gamma}}^{\mathcal{B},d}_{(\pi,r)} = \left\{(\boldsymbol{\gamma} \vee v_-) \wedge v_+,\ \boldsymbol{\gamma} \in \overline{\boldsymbol{S}}^{\mathcal{B},d}_{(\pi,r)}\right\}$ for $\overline{\boldsymbol{\Gamma}}^{\mathcal{B},d}_{(\boldsymbol{s},r)}$, where $\overline{\boldsymbol{S}}^{\mathcal{B},d}_{(\pi,r)}$ characterises the space of piecewise polynomial functions on $[0,1]^d$ such that on each hyperrectangle of π, any $\boldsymbol{\gamma} \in \overline{\boldsymbol{S}}^{\mathcal{B},d}_{(\pi,r)}$ 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 $\overline{\boldsymbol{\Gamma}}^{\mathcal{B},d}_{(\pi,r)}$ on $\mathscr{W}$ is given by

$$V_{(\pi,r)} = (r+1)^d |\pi| + 1, \tag{3.3.5}$$

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 (3.3.4) as

$$\Delta(\pi,r) = \log(8d)|\pi| + r. \tag{3.3.6}$$

3.3.2 Adaptivity result

Before deriving the risk bound for our estimator based on the constructed family in Section 3.3.1, we first discuss the parametrization issue of the exponential family. As it has been shown in Section 2.4.1 and 2.4.2 of Chapter 2, 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 2.4.2 that if we parametrize exponential families by their means, 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 3.3.1. The exponential family $\overline{\mathscr{Q}} = \{R_\gamma,\ \gamma \in J\}$ has been parametrized in the way that there exists a constant $\kappa > 0$ such that

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

Let us remark that, by Proposition 2.4.1, Assumption 3.3.1 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 by (2.2.1).

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 $\left\{\overline{\boldsymbol{\Gamma}}^{\mathcal{B},d}_{(\boldsymbol{s},r)},\ (\boldsymbol{s}, r) \in \mathcal{M}\right\}$ defined in Section 3.3.1, the associated countable subsets $\left\{\boldsymbol{\Gamma}^{\mathcal{B},d}_{(\boldsymbol{s},r)},\ (\boldsymbol{s}, r) \in \mathcal{M}\right\}$ and the weights defined by (3.3.4), we are now able to apply the model selection procedure introduced in Section 3.2.2 to estimate $\boldsymbol{\gamma}^\star$. The following result shows that under Assumption 3.3.1, the resulted estimator $\widehat{\boldsymbol{\gamma}}(\boldsymbol{X})$ based on $\left\{\boldsymbol{\Gamma}^{\mathcal{B},d}_{(\boldsymbol{s},r)},\ (\boldsymbol{s}, r) \in \mathcal{M}\right\}$ 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 3.3.1. *Under Assumption 3.3.1, whatever the distribution of W, the estimator $\widehat{\boldsymbol{\gamma}}(\boldsymbol{X})$ given by the model selection procedure in Section 3.2.2 over the countable family $\left\{\boldsymbol{\Gamma}^{\mathcal{B},d}_{(\boldsymbol{s},r)},\ (\boldsymbol{s}, r) \in \mathbb{N}^d \times \mathbb{N}\right\}$ with the weights defined by (3.3.4) 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 3.3.1 is postponed to Section 3.7.1. We hereby give some comments on this result. First, Corollary 3.3.1 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 shall 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 (2019), 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 proved in Proposition 2.4.3. 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 3.3.1, 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 3.3.1, where the numerical constant depends on K, κ, d, $\boldsymbol{\alpha}$ and p.

3.4 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 $Q_i^\star = R_{\boldsymbol{\gamma}^\star}$ for all $i \in \{1, \ldots, n\}$ 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 $\boldsymbol{w} = (w_1, \ldots, w_k) \in [0,1]^k$. For a vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k) \in (\mathbb{R}_+^*)^k$ with $\alpha_j = r_j + \alpha_j'$, $r_j \in \mathbb{N}$ and $\alpha_j' \in (0,1]$ for $j \in \{1, \ldots, k\}$, Hölder space $\mathcal{H}^{\boldsymbol{\alpha}}([0,1]^k)$ denotes the collection of functions f on $[0,1]^k$ satisfying for any $(w_1, \ldots, w_{j-1}, w_{j+1}, \ldots, w_k) \in [0,1]^{k-1}$ and all $x, y \in [0,1]$

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

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 $[v_-, v_+] \subset J$ with $v_- < v_+$.

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

$$I_j'(h_j) = \begin{cases} [0, 1/t_j] & , \quad h_j = 0, \\ (h_j/t_j, (h_j+1)/t_j] & , \quad h_j = 1, \ldots, t_j - 1. \end{cases} \tag{3.4.1}$$

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 3.3.1, the following result is easy to obtain.

Lemma 3.4.1. *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 $\|\cdot\|_\infty$.*

3.4.1 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 $\boldsymbol{\gamma} : [0,1]^d \to [v_-, v_+] \subset J$ 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 $B^{\beta_j}_{p_j,p_j}([0,1], R_j)$, we consider the family $\left\{\overline{\boldsymbol{S}}^{\mathcal{B},1}_{(s,r)},\ (s,r) \in \mathbb{N} \times \mathbb{N}\right\}$ introduced in Section 3.3.1 taking $d = 1$. We recall that the functions belonging to the above family are built based on the collection of particular partitions $M^{\mathcal{B},1} = \cup_{s\in\mathbb{N}} M^{\mathcal{B},1}_s$. Therefore, we can rewrite the family in an alternative way $\left\{\overline{\boldsymbol{S}}^{\mathcal{B},1}_{(\pi,r)},\ (\pi,r) \in M^{\mathcal{B},1} \times \mathbb{N}\right\}$. To approximate the Hölder class of functions $\mathcal{H}^\alpha([0,1], L)$ with values in $[v_-, v_+]$, we consider the family $\left\{\overline{\boldsymbol{\Gamma}}^{\mathcal{H},1}_{(t,r)},\ (t,r) \in \mathbb{N}^* \times \mathbb{N}\right\}$, where $\overline{\boldsymbol{\Gamma}}^{\mathcal{H},1}_{(t,r)} = \left\{(\boldsymbol{\gamma} \vee v_-) \wedge v_+,\ \boldsymbol{\gamma} \in \overline{\boldsymbol{S}}^{\mathcal{H},1}_{(t,r)}\right\}$.

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

$$\boldsymbol{\gamma}(\boldsymbol{w}) = f\left[(g(\boldsymbol{w}) \vee 0) \wedge 1\right], \quad \text{for all } \boldsymbol{w} = (w_1, \ldots, w_d) \in [0,1]^d, \tag{3.4.2}$$

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 3.4.1. *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*

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

where $U = t + r + 2$.

The proof is postponed to Section 3.7.3. 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 3.3.1, $\boldsymbol{\Gamma}^{\mathcal{H},1}_{(t,r)} = \left\{(\boldsymbol{\gamma} \vee v_-) \wedge v_+,\ \boldsymbol{S}^{\mathcal{H},1}_{(t,r)}\right\}$ 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

$$\Delta(\boldsymbol{\pi}, t, r) = 3\log 2\left(\sum_{j=1}^d |\pi_j|\right) + r + t. \tag{3.4.3}$$

The following result shows inequality (3.2.2) is satisfied with the weights defined by (3.4.3).

Lemma 3.4.2. *With the weights defined by (3.4.3), 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 3.4.1 and Lemma 3.4.2, we can apply the model selection procedure introduced in Section 3.2.2 and obtain the following.

Corollary 3.4.1. *Under Assumption 3.3.1, no matter what the distribution of W is, the estimator $\widehat{\boldsymbol{\gamma}}(\boldsymbol{X})$ given by the model selection procedure in Section 3.2.2 over the family $\left\{\boldsymbol{\Gamma}^A_{(\boldsymbol{\pi},t,r)},\ (\boldsymbol{\pi},t,r)\in(M^{\mathcal{B},1})^d\times\mathbb{N}^*\times\mathbb{N}\right\}$ with the weights defined by (3.4.3) satisfies for all $\alpha, L\in\mathbb{R}^*_+$ and $\boldsymbol{\beta},\mathbf{p},\mathbf{R}\in(\mathbb{R}^*_+)^d$ such that $\beta_j>1/p_j$*

$$\begin{aligned}&\sup_{\boldsymbol{\gamma}^\star\in\mathcal{F}_{[v_-,v_+]}(\alpha,\boldsymbol{\beta},\mathbf{p},L,\mathbf{R})} C'_{\kappa,d,\alpha,\boldsymbol{\beta},\mathbf{p}}\mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star},R_{\widehat{\boldsymbol{\gamma}}})\right]\\ &\leq\left\{\left[\sum_{j=1}^d\left(LR_j^{\alpha\wedge1}\right)^{\frac{2}{2(\alpha\wedge1)\beta_j+1}}n^{-\frac{2(\alpha\wedge1)\beta_j}{2(\alpha\wedge1)\beta_j+1}}\right]+L^{\frac{2}{2\alpha+1}}n^{-\frac{2\alpha}{2\alpha+1}}+\frac{1}{n}\right\}\mathcal{L}_n^2,\end{aligned}\tag{3.4.4}$$

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 3.4.1 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 $Q^\star\neq R_{\boldsymbol{\gamma}^\star}$ 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 3.4.1. However, as long as the bias term is not too large compared to the quantity on the right hand side of (3.4.4), 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.

3.4.2 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

$$\boldsymbol{\gamma}(\boldsymbol{w})=f\circ g(\boldsymbol{w}),\quad\text{for all }\boldsymbol{w}=(w_1,\ldots,w_d)\in[0,1]^d,\tag{3.4.5}$$

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

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

and $f \in \mathcal{H}^{\boldsymbol{\alpha}}(L, v_-, v_+)$ mapping $[0,1]^l$ to $[v_-, v_+] \subset J$ 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 $\left\{\overline{\boldsymbol{\Gamma}}_{(\boldsymbol{t},r)}^{\mathcal{H},l},\ (\boldsymbol{t}, r) \in (\mathbb{N}^*)^l \times \mathbb{N}\right\}$, where $\overline{\boldsymbol{\Gamma}}_{(\boldsymbol{t},r)}^{\mathcal{H},l}$ stands for the set $\left\{(\boldsymbol{\gamma} \vee v_-) \wedge v_+,\ \boldsymbol{\gamma} \in \overline{\boldsymbol{S}}_{(\boldsymbol{t},r)}^{\mathcal{H},l}\right\}$. 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 $\overline{\boldsymbol{\Gamma}}_{(\boldsymbol{t},r)}^{M}$ on $\mathscr{W} = [0,1]^d$ as

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

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

Proposition 3.4.2. *For any $r \in \mathbb{N}$ and $\boldsymbol{t} = (t_1, \ldots, t_l) \in (\mathbb{N}^*)^l$, the class of functions $\overline{\boldsymbol{\Gamma}}_{(\boldsymbol{t},r)}^{M}$ is a VC-subgraph on $\mathscr{W} = [0,1]^d$ with dimension*

$$V_{(\boldsymbol{t},r)}^{M} \leq 2 + \left\lfloor 2ld + \left(\prod_{j=1}^{l} t_j\right)(r+1)^l \right\rfloor \log_2\left[4eU \log_2(2eU)\right], \tag{3.4.6}$$

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

The proof is postponed to Section 3.7.3. For any $r \in \mathbb{N}$ and $\boldsymbol{t} = (t_1, \ldots, t_l) \in (\mathbb{N}^*)^l$, we take the countable subset $\boldsymbol{\Gamma}_{(\boldsymbol{t},r)}^{M}$ of $\overline{\boldsymbol{\Gamma}}_{(\boldsymbol{t},r)}^{M}$ defined as

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

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

$$\Delta(\boldsymbol{t}, r) = \sum_{j=1}^{l} t_j + r. \tag{3.4.7}$$

The following result shows inequality (3.2.2) is satisfied with the weights defined by (3.4.7).

Lemma 3.4.3. *With the weights defined by (3.4.7), 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 Section 3.7.2. With Proposition 3.4.2 and Lemma 3.4.3, we are able to apply the model selection procedure introduced in Section 3.2.2 and obtain the following.

Corollary 3.4.2. *Under Assumption 3.3.1, no matter what the distribution of W is, the estimator $\widehat{\boldsymbol{\gamma}}(\boldsymbol{X})$ given by the model selection procedure in Section 3.2.2 over the family $\left\{\boldsymbol{\Gamma}^M_{(\boldsymbol{t},r)},\ (\boldsymbol{t},r)\in(\mathbb{N}^*)^l\times\mathbb{N}\right\}$ with the weights defined by (3.4.7) 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 3.4.2. But as long as the conditional distribution $Q^\star$ 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 (3.4.5) 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 3.4.1 and 3.4.2.

3.5 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 $Q_i^\star=R_{\boldsymbol{\gamma}^\star}$ 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

$$f : \mathbb{R}^d \to \mathbb{R}, \quad \boldsymbol{w} \mapsto f(\boldsymbol{w}) = M_L \circ \sigma \circ M_{L-1} \circ \cdots \circ \sigma \circ M_0(\boldsymbol{w}), \tag{3.5.1}$$

where

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

A_l is a $p \times p$ weight matrix for $l \in \{1, \ldots, L-1\}$, 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 $l \in \{0, \ldots, L-1\}$, 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 $l \in \{0, \ldots, L\}$, 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 $l \in \{1, \ldots, L-1\}$, $\boldsymbol{s}_l$ is of size $p(p+1)$ and for $l = L$, $\boldsymbol{s}_L$ is of size $p+1$. Essentially, indicator vectors $\boldsymbol{s}_l$, $l \in \{0, \ldots, L\}$ represent collections of functions based on the structure of neural networks. The last p components in $\boldsymbol{s}_l$, $l \in \{0, \ldots, L-1\}$ 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 $\boldsymbol{s} = (\boldsymbol{s}_0^\top, \ldots, \boldsymbol{s}_L^\top)^\top$, 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 3.5.1. *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 Section 3.7.3. 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 3.5.1 entails the VC dimension of $\overline{\boldsymbol{S}}_{(L,p)}$ is, up to a constant, bounded by $\overline{p}L \log\left[(L+1)\left(pL/2+1\right)\right]$.

3.5.1 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 $[v_-, v_+] \subset J$. 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 $[v_-, v_+] \subset J$ and is of the form

$$f(w) = \sum_{k \in \mathbb{N}^*} t^k g\left(h^{\circ k}(w)\right), \quad \text{for all } w \in [0,1], \tag{3.5.2}$$

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

$$h(w) = \begin{cases} 2w & , \quad 0 \leq w \leq \dfrac{1}{2}, \\ 2(1-w) & , \quad \dfrac{1}{2} < w \leq 1. \end{cases} \tag{3.5.3}$$

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 3.2.1 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

$$\Delta(L,p) = L + p. \tag{3.5.4}$$

As an immediate consequence, we have $\Sigma = \sum_{(L,p) \in (\mathbb{N}^*)^2} e^{-\Delta(L,p)} \leq 1$ which satisfies the inequality (3.2.2). Therefore, we are able to apply the model selection procedure introduced in Section 3.2.2 and obtain the following result.

Corollary 3.5.1. *Let Assumption 3.3.1 hold true. Whatever the distribution of W, the estimator $\widehat{\boldsymbol{\gamma}}(\boldsymbol{X})$ given by the model selection procedure in Section 3.2.2 over the family $\left\{ \boldsymbol{\Gamma}_{(L,p)},\ (L,p) \in \mathbb{N}^* \times \mathbb{N}^* \right\}$ with the weights defined by (3.5.4) 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 $\overline{\mathscr{Q}} = \{R_\gamma,\ \gamma \in J\}$ by taking $\gamma = \theta/(2\sqrt{2}\sigma)$, where θ is the mean so that according to Proposition 2.4.1, Assumption 3.3.1 is satisfied with $\kappa = 1$ and $J = \mathbb{R}$. We therefore can take v_- the smallest integer in computer and v_+ the largest so that $[v_-, v_+] \subset J$. Let $\boldsymbol{\gamma}^\star = \sum_{k \in \mathbb{N}^*} 2^{-k} h^{\circ k}$ with

h defined by (3.5.3) 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 (3.5.2). 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 3.5.1, the estimator $\widehat{\boldsymbol{\gamma}}(\boldsymbol{X})$ obtained by the model selection procedure introduced in Section 3.2.2 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 in a miserably slow rate considering the minimax converge rate for an α-smooth function is of order $n^{-2\alpha/(2\alpha+1)}$.

3.5.2 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

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

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

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

tions with values in $[v_-, v_+] \subset J$ as,

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

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 $(L, p) \in (\mathbb{N}^*)^2$, $\boldsymbol{s} = (\boldsymbol{s}_0^\top, \ldots, \boldsymbol{s}_L^\top)^\top \in \{0,1\}^{\overline{p}}$ 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 3.5.1. 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

$$\Delta(L, p, \boldsymbol{s}) = \begin{cases} \|\boldsymbol{s}\|_0 \log\left(\dfrac{2e\overline{p}}{\|\boldsymbol{s}\|_0}\right) + p + L & ,\quad \|\boldsymbol{s}\|_0 \neq 0, \\ p + L & ,\quad \|\boldsymbol{s}\|_0 = 0. \end{cases} \tag{3.5.5}$$

The following result shows (3.2.2) is satisfied with the associated weights defined by (3.5.5).

Lemma 3.5.1. *For any $L \in \mathbb{N}^*$, $p \in \mathbb{N}^*$ and $\boldsymbol{s} = (\boldsymbol{s}_0^\top, \ldots, \boldsymbol{s}_L^\top)^\top \in \{0,1\}^{\overline{p}}$, we define $\Delta(L, p, \boldsymbol{s})$ by (3.5.5). Then,*

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

For any $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_k) \in (\mathbb{R}_+^*)^{k+1}$, we define the effective smoothness indices by $\alpha_i' = \alpha_i \prod_{l=i+1}^k (\alpha_l \wedge 1)$ for all $i \in \{0, \ldots, k-1\}$ and $\alpha_k' = \alpha_k$. We denote $\phi_n = \max_{i=0,\ldots,k} n^{-2\alpha_i'/(2\alpha_i' + t_i)}$. Combining the result of Lemma 3.5.1 and Proposition 3.5.1, we are now able to apply the model selection procedure in Section 3.2.2. 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 3.5.2. *Let Assumption 3.3.1 hold true. Whatever the distribution of W, the estimator $\widehat{\boldsymbol{\gamma}}(\boldsymbol{X})$ given by the model selection procedure in Section 3.2.2 over the family $\big\{\boldsymbol{\Gamma}_{(L,p,\boldsymbol{s})},\ (L, p, \boldsymbol{s}) \in (\mathbb{N}^*)^2 \times \{0,1\}^{\overline{p}}\big\}$ with the weights defined by (3.5.5) satisfies with a sufficiently large n, 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 $j \in \{0, \ldots, k\}$ and $\boldsymbol{\alpha} \in (\mathbb{R}_+^*)^{k+1}$,*

$$\sup_{\boldsymbol{\gamma}^\star \in \mathcal{F}_{[v_-,v_+]}(k,\mathbf{d},\mathbf{t},\boldsymbol{\alpha},K)} \mathbb{E}\left[h^2\left(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}}\right)\right] \le C_{\kappa,k,\mathbf{d},\mathbf{t},\boldsymbol{\alpha},K} \phi_n \log^4 n, \tag{3.5.6}$$

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 3.5.2, 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 the MLE. When there is a misspecification or data contamination, as long as the bias remains small compared to the right hand side of (3.5.6), the behaviour of our estimator will be of the same order as the model is exact.

3.6 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 integer p stands for the number of the explanatory variables. This number may be large, possibly larger than n. The exponential family $\overline{\mathscr{Q}} = \{R_\gamma = r_\gamma \cdot \nu,\ \gamma \in J\}$ is parametrized in its natural form, i.e. for all $y \in \mathscr{Y}$, $\gamma \in J$,

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

which is the particular situation when taking u as the identity function in (3.1.2). We assume that there exists an unknown function $\boldsymbol{\gamma}^\star$ on $\mathscr{W}$ taking values in $[v_-, v_+] \subset J$ with $v_- < v_+$ as a linear combination of some subset of the p explanatory variables, namely

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

with $\gamma_j^\star \in \mathbb{R}$, such that the conditional distribution of Y_i given $W_i = w_i$ 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 $\Omega = \{1, \dots, p\}$ 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

$$\boldsymbol{\gamma}(\boldsymbol{w}) = \sum_{j=1}^{p} \gamma_j w^{(j)} \quad \text{for all } \boldsymbol{w} \in \mathscr{W}, \tag{3.6.1}$$

where the coordinates of $\widetilde{\gamma} = (\gamma_1, \dots, \gamma_p) \in \mathbb{R}^p$ 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 (3.6.1) 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 $\overline{\mathbf{\Gamma}}_m = \left\{(\boldsymbol{\gamma} \vee v_-) \wedge v_+,\ \boldsymbol{\gamma} \in \overline{\boldsymbol{S}}_m\right\}$ and $\mathbf{\Gamma}_m = \{(\boldsymbol{\gamma} \vee v_-) \wedge v_+,\ \boldsymbol{\gamma} \in \boldsymbol{S}_m\}$. 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 $\mathbf{\Gamma}_m$ in $\overline{\mathbf{\Gamma}}_m$, $m \in \mathcal{M}$.

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

$$\Delta(m) = \begin{cases} 2\log(1 + |m|) & ,\quad m \in \mathcal{M}_o, \\ |m| \log\left(\dfrac{2ep}{|m|}\right) & ,\quad m \in \mathcal{M} \backslash \mathcal{M}_o. \end{cases} \tag{3.6.2}$$

The following result shows with the weights defined by (3.6.2), inequality (3.2.2) is satisfied.

Lemma 3.6.1. *Let $\mathcal{M} = \mathcal{P}(\Omega)$. For any $m \in \mathcal{M}$, the weight is defined by (3.6.2). Then $\Sigma = \sum_{m \in \mathcal{M}} e^{-\Delta(m)} \leq 1 + \pi^2/6$.*

Moreover, for any $m \in \mathcal{M}$, $\overline{\boldsymbol{S}}_m$ defined by (3.6.1) is a $|m|$-dimensional vector space. As an immediate consequence of Proposition 1.5.1, $\overline{\mathbf{\Gamma}}_m$ is VC-subgraph on $\mathscr{W}$ with dimension not larger than $|m| + 1$ which satisfies the Assumption 3.2.1 with $V_m = |m| + 1$. We are now able to apply the model selection procedure presented in Section 3.2.2 and obtain the following result.

Corollary 3.6.1. *For all $m \in \mathcal{M}$, let $\overline{\mathscr{Q}}_m = \left\{R_{\boldsymbol{\gamma}},\ \boldsymbol{\gamma} \in \overline{\mathbf{\Gamma}}_m\right\}$. Whatever the distribution of W, the estimator $R_{\widehat{\boldsymbol{\gamma}}}$ given by the model selection procedure in Section 3.2.2 over $\{\mathbf{\Gamma}_m,\ m \in \mathcal{M}\}$ associated with the weight defined by (3.6.2) satisfies*

$$\mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}})\right] \leq 1.95 \times 10^7 (\mathscr{B}_o \wedge \mathscr{B}_c), \tag{3.6.3}$$

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 3.6.1 is postponed to Section 3.7.1. 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

$$\mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}})\right] \leq 1.95 \times 10^7 \mathscr{B}_c. \tag{3.6.4}$$

Comparing (3.6.4) with the one given in (3.6.3), we note that (3.6.3) improves it by a $\log(p)$ term whenever the minimizer $m^\star \in \mathcal{M}$ in the right hand side of (3.6.3) does belong to $\mathcal{M}_o$.

3.7 Proofs

3.7.1 Proofs of the main theorem and its corollaries

Proof of Theorem 3.2.1

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

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

Moreover, according to (1.4.1), we have for any $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}$,

$$\begin{aligned} \mathbf{h}^2(\mathbf{Q}^\star, \mathbf{R}_{\boldsymbol{\gamma}}) &= \sum_{i=1}^n \int_{\mathscr{W}} h^2\left(Q_i^\star(w), R_{\boldsymbol{\gamma}(w)}\right) dP_{W_i}(w) \\ &= \sum_{i=1}^n h^2\left(P_i^\star, P_{i,\boldsymbol{\gamma}}\right) = \mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\boldsymbol{\gamma}}). \end{aligned} \tag{3.7.1}$$

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

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

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

Proposition 3.7.1. *Under Assumption 3.2.1, for any $m \in \mathcal{M}$, the class of functions $\mathcal{Q}_m = \{r_{\boldsymbol{\gamma}} :\ (w, y) \mapsto r_{\boldsymbol{\gamma}(w)}(y),\ \boldsymbol{\gamma} \in \overline{\boldsymbol{\Gamma}}_m\}$ on $\mathscr{X} = \mathscr{W} \times \mathscr{Y}$ is VC-subgragh with dimension not larger than $9.41 V_m$.*

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

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

where $A(\theta) = \log\left[\int_{\mathscr{Y}} \exp(\theta S(y)) d\nu(y)\right]$ and $\overline{\boldsymbol{\Theta}}_m = \{\boldsymbol{\theta} = u \circ \boldsymbol{\gamma},\ \boldsymbol{\gamma} \in \overline{\boldsymbol{\Gamma}}_m\}$. By Proposition 1.5.2 (Proposition 42 of Baraud et al. (2017)), VC-subgraph is preserved by composition with a monotone function. Therefore, under Assumption 3.2.1, $\overline{\boldsymbol{\Theta}}_m$ is also VC-subgraph on $\mathscr{W}$ with dimension not larger than $V_m \geq 1$. Applying Proposition 2.7.1 with $\mathcal{Q} = \mathcal{Q}_m$ for each $m \in \mathcal{M}$, we can conclude. □

The next result provides an upper bound for the ρ-dimension function $D^{\mathscr{P}_m}$ of $\mathscr{P}_m$.

Proposition 3.7.2. *Under Assumption 3.2.1, 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 $\overline{P}_i = \overline{p} \cdot \mu_i$ for all $i \in \{1, \ldots, n\}$,*

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

Proof. The proof is basically similar to the proof of Proposition 2.7.2 except a modification of the class $\mathscr{F}_y$. More precisely, for any $y > 0$, we define

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

Then combining Proposition 3.7.1, the conclusion is easy to obtain by following the proof of Proposition 2.7.2. □

Now we turn to prove Theorem 3.2.1. It follows by Proposition 3.7.2 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 $\mathbf{P}_{\widehat{\boldsymbol{\gamma}}}$ satisfies, for all $\xi > 0$ with a probability at least $1 - \Sigma e^{-\xi}$,

$$\mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\widehat{\boldsymbol{\gamma}}}) \leq \inf_{m \in \mathcal{M}} \left[c_1 \mathbf{h}^2(\mathbf{P}^\star, \mathscr{P}_m) + c_2 \left(\Xi(m) + 1.49 + \xi \right) \right], \tag{3.7.2}$$

where $c_1 = 150$ and $c_2 = 5014$. 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 $\mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\widehat{\boldsymbol{\gamma}}}) = \mathbf{h}^2(\mathbf{Q}^\star, \mathbf{R}_{\widehat{\boldsymbol{\gamma}}})$ and $\mathbf{h}^2(\mathbf{P}^\star, \mathscr{P}_m) = \mathbf{h}^2(\mathbf{Q}^\star, \mathscr{Q}_m)$, for all $m \in \mathcal{M}$.

Proof of Corollary 3.3.1

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

Proposition 3.7.3. *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 $f \in B^{\boldsymbol{\alpha}}_{p,q}\left([0,1]^d, R\right)$ *and all* $l \in \mathbb{N}$, *there exists a partition* $\pi(l) \in \cup_{\boldsymbol{s}\in\mathbb{N}^d} M^{\mathcal{B},d}_{\boldsymbol{s}}$ *of* $[0,1]^d$ *containing only hyperrectangles such that*

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

and

$$\inf_{\widetilde{f}\in\overline{\boldsymbol{S}}^{\mathcal{B},d}_{(\pi(l),r)}} \|f - \widetilde{f}\|_{p'} \leq C_{d,r,\boldsymbol{\alpha},p,p'} R 2^{-l\overline{\alpha}}, \tag{3.7.3}$$

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 3.3.1. Under Assumption 3.3.1, applying (3.2.8), Lemma 3.3.1 and 3.3.2, 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 $\boldsymbol{\gamma}^\star \in B^{\boldsymbol{\alpha}}_{p,q}(R, v_-, v_+)$

$$\begin{aligned}
&\mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}})\right] \\
\leq& c_2\left(c_3 + \frac{e}{e-1}\right) \inf_{(\boldsymbol{s},r)\in\mathcal{M}} \left[h^2(R_{\boldsymbol{\gamma}^\star}, \overline{\mathscr{Q}}^d_{(\boldsymbol{s},r)}) + \frac{\Delta(\boldsymbol{s},r)}{n} + \frac{V_{(\boldsymbol{s},r)}}{n}(1+\log n)\right] \\
\leq& C_\kappa \inf_{(\boldsymbol{s},r)\in\mathcal{M}} \left[\inf_{\overline{\boldsymbol{\gamma}}\in\overline{\boldsymbol{\Gamma}}^{\mathcal{B},d}_{(\boldsymbol{s},r)}} \|\boldsymbol{\gamma}^\star - \overline{\boldsymbol{\gamma}}\|^2_{2,P_W} + \frac{\Delta(\boldsymbol{s},r)}{n} + \frac{V_{(\boldsymbol{s},r)}}{n}(1+\log n)\right],
\end{aligned} \tag{3.7.4}$$

where $\overline{\mathscr{Q}}^d_{(\boldsymbol{s},r)} = \left\{R_{\boldsymbol{\gamma}},\ \boldsymbol{\gamma} \in \overline{\boldsymbol{\Gamma}}^{\mathcal{B},d}_{(\boldsymbol{s},r)}\right\}$ and C_κ is a constant depending on κ only. We then apply Proposition 3.7.3 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

$$\Delta(\pi(l), r) = \log(8d)|\pi(l)| + r \leq C_{d,\boldsymbol{\alpha},p} 2^{ld}, \tag{3.7.5}$$

$$V_{(\pi(l),r)} = (r+1)^d |\pi(l)| + 1 \leq C_{d,\boldsymbol{\alpha},p} 2^{ld}, \tag{3.7.6}$$

$$\begin{aligned}
\inf_{\overline{\boldsymbol{\gamma}}\in\overline{\boldsymbol{\Gamma}}^{\mathcal{B},d}_{(\pi(l),r)}} \|\boldsymbol{\gamma}^\star - \overline{\boldsymbol{\gamma}}\|^2_{2,P_W} &= \inf_{\overline{\boldsymbol{\gamma}}\in\overline{\boldsymbol{\Gamma}}^{\mathcal{B},d}_{(\pi(l),r)}} \int_{\mathscr{W}} |\boldsymbol{\gamma}^\star(w) - \overline{\boldsymbol{\gamma}}(w)|^2 dP_W(w) \\
&\leq \inf_{\overline{\boldsymbol{\gamma}}\in\overline{\boldsymbol{\Gamma}}^{\mathcal{B},d}_{(\pi(l),r)}} \|\boldsymbol{\gamma}^\star - \overline{\boldsymbol{\gamma}}\|^2_\infty \\
&\leq \inf_{\overline{\boldsymbol{\gamma}}\in\overline{\boldsymbol{S}}^{\mathcal{B},d}_{(\pi(l),r)}} \|\boldsymbol{\gamma}^\star - \overline{\boldsymbol{\gamma}}\|^2_\infty \\
&\leq C_{d,\boldsymbol{\alpha},p} R^2 2^{-2l\overline{\alpha}}.
\end{aligned} \tag{3.7.7}$$

Plugging (3.7.5), (3.7.6) and (3.7.7) into (3.7.4), we derive

$$\begin{aligned}
&\mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}})\right] \\
\leq & C_\kappa \inf_{l\in\mathbb{N}} \left[\inf_{\overline{\boldsymbol{\gamma}}\in\overline{\boldsymbol{\Gamma}}^{\mathcal{B},d}_{(\pi(l),r)}} \|\boldsymbol{\gamma}^\star - \overline{\boldsymbol{\gamma}}\|^2_{2,P_W} + \frac{\Delta(\pi(l),r)}{n} + \frac{V_{(\pi(l),r)}}{n}(1+\log n) \right] \\
\leq & C_{\kappa,d,\boldsymbol{\alpha},p} \inf_{l\in\mathbb{N}} \left(R^2 2^{-2l\overline{\alpha}} + \frac{2^{ld}}{n} \right)(1+\log n), \qquad (3.7.8)
\end{aligned}$$

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 (3.7.8). If $nR^2 < 1$, we take $l = 0$ so that

$$R^2 2^{-2l\overline{\alpha}} + \frac{2^{ld}}{n} = R^2 + \frac{1}{n} < \frac{2}{n}. \qquad (3.7.9)$$

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,

$$R^2 2^{-2l\overline{\alpha}} + \frac{2^{ld}}{n} \leq 2R^2 2^{-2l\overline{\alpha}} \leq C_{\boldsymbol{\alpha}} R^{\frac{2d}{d+2\overline{\alpha}}} n^{-\frac{2\overline{\alpha}}{d+2\overline{\alpha}}}, \qquad (3.7.10)$$

where $C_{\boldsymbol{\alpha}}$ is a constant depending only on $\boldsymbol{\alpha}$. Combining (3.7.8), (3.7.9) and (3.7.10), 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 $B^{\boldsymbol{\alpha}}_{p,q}(R, v_-, v_+)$.

Proof of Corollary 3.4.1

Lemma 3.7.1. *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 3.7.2. *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 3.7.4. *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*

$$\inf_{\widetilde{f} \in \overline{\boldsymbol{S}}^{\mathcal{H},k}_{(\boldsymbol{t},r)}} \|f - \widetilde{f}\|_\infty \leq C_{k,r} L \sum_{j=1}^k t_j^{-\alpha_j}, \tag{3.7.11}$$

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

Now we turn to prove Corollary 3.4.1. 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 $[f\,[(g \vee 0) \wedge 1] \vee v_-] \wedge v_+$, 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

$$\begin{aligned}
&\sup_{\boldsymbol{w}\in[0,1]^d} \left| \boldsymbol{\gamma}\left(\sum_{j=1}^d \boldsymbol{\gamma}'_j(w_j)\right) - \left[f\left[\left(\left(\sum_{j=1}^d g_j(w_j)\right) \vee 0\right) \wedge 1\right] \vee v_- \right] \wedge v_+ \right| \\
\leq &\sup_{\boldsymbol{w}\in[0,1]^d} \left| \boldsymbol{\gamma}\left(\sum_{j=1}^d \boldsymbol{\gamma}'_j(w_j)\right) - f\left[\left(\left(\sum_{j=1}^d g_j(w_j)\right) \vee 0\right) \wedge 1\right] \right| \\
\leq &\sup_{\boldsymbol{w}\in[0,1]^d} \left| \boldsymbol{\gamma}\left(\sum_{j=1}^d \boldsymbol{\gamma}'_j(w_j)\right) - \boldsymbol{\gamma}\left[\left(\left(\sum_{j=1}^d g_j(w_j)\right) \vee 0\right) \wedge 1\right] \right| + \|\boldsymbol{\gamma} - f\|_\infty \\
\leq &L \sup_{\boldsymbol{w}\in[0,1]^d} \left| \left(\sum_{j=1}^d \boldsymbol{\gamma}'_j(w_j)\right) - \left(\sum_{j=1}^d g_j(w_j)\right) \right|^{\alpha\wedge 1} + \|\boldsymbol{\gamma} - f\|_\infty \\
\leq &L \left\| \left(\sum_{j=1}^d |\boldsymbol{\gamma}'_j - g_j|\right)^{\alpha\wedge 1} \right\|_\infty + \|\boldsymbol{\gamma} - f\|_\infty .
\end{aligned} \tag{3.7.12}$$

We then apply Lemma 3.7.1 and Lemma 3.7.2 with $k=\infty$, μ being the Lebesgue measure (probability) and $u(z)=z^{\alpha\wedge 1}$ to (3.7.12) and obtain

$$\begin{aligned}
&\left\|\boldsymbol{\gamma}\left(\sum_{j=1}^{d}\boldsymbol{\gamma}_j'\right)-\left[f\left[(g\vee 0)\wedge 1\right]\vee v_-\right]\wedge v_+\right\|_\infty\\
\leq& L\sum_{j=1}^{d}\left\||\boldsymbol{\gamma}_j'-g_j|^{\alpha\wedge 1}\right\|_\infty+\|\boldsymbol{\gamma}-f\|_\infty\\
\leq& L\sum_{j=1}^{d}\left(\left\|\boldsymbol{\gamma}_j'-g_j\right\|_\infty\right)^{\alpha\wedge 1}+\|\boldsymbol{\gamma}-f\|_\infty\,. \qquad (3.7.13)
\end{aligned}$$

We take

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

By Proposition 3.7.3, 3.7.4 and Lemma 3.3.1, 3.4.1, 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 $(\boldsymbol{l},t)=(l_1,\dots,l_d,t)\in\mathbb{N}^d\times\mathbb{N}^*$ 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

$$\inf_{f\in\boldsymbol{S}^{\mathcal{H},1}_{(t,r)}}\|\boldsymbol{\gamma}-f\|_\infty=\inf_{f\in\overline{\boldsymbol{S}}^{\mathcal{H},1}_{(t,r)}}\|\boldsymbol{\gamma}-f\|_\infty\leq C_{\alpha,\boldsymbol{\beta}}Lt^{-\alpha} \qquad (3.7.14)$$

and

$$\inf_{g_j\in\boldsymbol{S}^{\mathcal{B},1}_{(\pi(l_j),r)}}\|\boldsymbol{\gamma}_j'-g_j\|_\infty=\inf_{g_j\in\overline{\boldsymbol{S}}^{\mathcal{B},1}_{(\pi(l_j),r)}}\|\boldsymbol{\gamma}_j'-g_j\|_\infty\leq C_{\alpha,\boldsymbol{\beta},p_j}R_j2^{-l_j\beta_j}. \qquad (3.7.15)$$

Combining (3.7.13), (3.7.14) and (3.7.15), 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 $(\boldsymbol{l},t)=(l_1,\dots,l_d,t)\in\mathbb{N}^d\times\mathbb{N}^*$ and any $\boldsymbol{\gamma}\left(\sum_{j=1}^{d}\boldsymbol{\gamma}_j'\right)\in\mathcal{F}_{[v_-,v_+]}(\alpha,\boldsymbol{\beta},\mathbf{p},L,\mathbf{R})$,

$$\begin{aligned}
&\inf_{f\in\boldsymbol{S}^{\mathcal{H},1}_{(t,r)},\,g_j\in\boldsymbol{S}^{\mathcal{B},1}_{(\pi(l_j),r)}}\left\|\boldsymbol{\gamma}\left(\sum_{j=1}^{d}\boldsymbol{\gamma}_j'\right)-\left[f\left[\left(\left(\sum_{j=1}^{d}g_j\right)\vee 0\right)\wedge 1\right]\vee v_-\right]\wedge v_+\right\|_{2,P_W}^2\\
\leq&\inf_{f\in\boldsymbol{S}^{\mathcal{H},1}_{(t,r)},\,g_j\in\boldsymbol{S}^{\mathcal{B},1}_{(\pi(l_j),r)}}\left\|\boldsymbol{\gamma}\left(\sum_{j=1}^{d}\boldsymbol{\gamma}_j'\right)-\left[f\left[\left(\left(\sum_{j=1}^{d}g_j\right)\vee 0\right)\wedge 1\right]\vee v_-\right]\wedge v_+\right\|_\infty^2\\
\leq& C_d\left\{L^2\sum_{j=1}^{d}\left[\left(\inf_{g_j\in\boldsymbol{S}^{\mathcal{B},1}_{(\pi(l_j),r)}}\|\boldsymbol{\gamma}_j'-g_j'\|_\infty\right)^{\alpha\wedge 1}\right]^2+\inf_{f\in\boldsymbol{S}^{\mathcal{H},1}_{(t,r)}}\|\boldsymbol{\gamma}-f\|_\infty^2\right\}\\
\leq& C_{d,\alpha,\boldsymbol{\beta},\mathbf{p}}L^2\left[\sum_{j=1}^{d}R_j^{2(\alpha\wedge 1)}2^{-2(\alpha\wedge 1)l_j\beta_j}+t^{-2\alpha}\right]. \qquad (3.7.16)
\end{aligned}$$

We denote $\boldsymbol{\pi}(\boldsymbol{l}) = (\pi(l_1), \dots, \pi(l_d))$. For any $(l_1, \dots, l_d, t, r) \in \mathbb{N}^d \times \mathbb{N}^* \times \mathbb{N}$, by Proposition 3.4.1 and 3.7.3, we have

$$\begin{aligned} V^A_{(\boldsymbol{\pi}(\boldsymbol{l}),t,r)} + \Delta(\boldsymbol{\pi}(\boldsymbol{l}), t, r) \leq & C \left[\left(t + \sum_{j=1}^{d} |\pi(l_j)| \right) (r+1) \right] \log(t+r+2) \\ & + \left[(3 \log 2) \left(\sum_{j=1}^{d} |\pi(l_j)| \right) + r + t \right] \\ \leq & C_r \left(t + \sum_{j=1}^{d} |\pi(l_j)| \right) \log(t+r+2) \\ \leq & C_{\alpha,\boldsymbol{\beta},\mathbf{p}} \left(t + \sum_{j=1}^{d} 2^{l_j} \right) \log(t+r+2), \end{aligned} \tag{3.7.17}$$

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 3.3.1, applying (3.2.8) together with (3.7.16) and (3.7.17), 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 $(\boldsymbol{l}, t) = (l_1, \dots, l_d, t) \in \mathbb{N}^d \times \mathbb{N}^*$ and any $\boldsymbol{\gamma}\left(\sum_{j=1}^d \boldsymbol{\gamma}'_j\right) \in \mathcal{F}_{[v_-,v_+]}(\alpha, \boldsymbol{\beta}, \mathbf{p}, L, \mathbf{R})$,

$$\begin{aligned} \mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}})\right] \leq & C_\kappa \inf_{(\boldsymbol{\pi},t,r) \in (M^{\mathcal{B},1})^d \times (\mathbb{N}^*)^2} \left[\inf_{\widetilde{\boldsymbol{\gamma}} \in \boldsymbol{\Gamma}^A_{(\boldsymbol{\pi},t,r)}} \left\| \boldsymbol{\gamma}\left(\sum_{j=1}^d \boldsymbol{\gamma}'_j\right) - \widetilde{\boldsymbol{\gamma}} \right\|^2_{2,P_W} \right. \\ & \left. + \frac{\Delta(\boldsymbol{\pi}, t, r)}{n} + \frac{V_{(\boldsymbol{\pi},t,r)}}{n} (1 + \log n) \right] \\ \leq & C_{\kappa,d,\alpha,\boldsymbol{\beta},\mathbf{p}} (1 + \log n) \inf_{(l_1,\dots,l_d,t) \in \mathbb{N}^d \times \mathbb{N}^*} \left[\left(L^2 t^{-2\alpha} + \frac{t}{n} \right) \right. \\ & \left. + \sum_{j=1}^{d} \left(L^2 R_j^{2(\alpha \wedge 1)} 2^{-2(\alpha \wedge 1) l_j \beta_j} + \frac{2^{l_j}}{n} \right) \right] \log(t+r+2). \end{aligned} \tag{3.7.18}$$

To conclude, we need to optimize the right hand side of (3.7.18). 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

$$L^2 t^{-2\alpha} + \frac{t}{n} \leq 2\frac{t}{n} \leq \frac{2}{n} + 2L^{\frac{2}{2\alpha+1}} n^{-\frac{2\alpha}{2\alpha+1}}. \tag{3.7.19}$$

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

$$\log(t+r+2) \leq \log(r+3) = C_{\alpha,\boldsymbol{\beta}}. \tag{3.7.20}$$

Otherwise $nL^2 \geq 1$,

$$\begin{aligned}\log(t+r+2) &\leq \log\left[\left(nL^2\right)^{\frac{1}{2\alpha+1}} + r + 3\right] \\ &\leq \log\left[C_{\alpha,\boldsymbol{\beta}}\left(nL^2\right)^{\frac{1}{2\alpha+1}}\right] \\ &\leq C_{\alpha,\boldsymbol{\beta}}\left(\log n \vee \log L^2 \vee 1\right). \end{aligned} \tag{3.7.21}$$

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

$$L^2R_j^{2(\alpha\wedge 1)}2^{-2(\alpha\wedge 1)l_j\beta_j} + \frac{2^{l_j}}{n} < \frac{2}{n}. \tag{3.7.22}$$

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

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

which yields

$$\begin{aligned} L^2R_j^{2(\alpha\wedge 1)}2^{-2(\alpha\wedge 1)l_j\beta_j} + \frac{2^{l_j}}{n} &\leq L^2R_j^{2(\alpha\wedge 1)}2^{1-2(\alpha\wedge 1)l_j\beta_j} \\ &\leq C_{\alpha,\boldsymbol{\beta}}\left[L^2R_j^{2(\alpha\wedge 1)}\right]^{\frac{1}{2(\alpha\wedge 1)\beta_j+1}} n^{-\frac{2(\alpha\wedge 1)\beta_j}{2(\alpha\wedge 1)\beta_j+1}} \\ &\leq C_{\alpha,\boldsymbol{\beta}}\left(LR_j^{\alpha\wedge 1}\right)^{\frac{2}{2(\alpha\wedge 1)\beta_j+1}} n^{-\frac{2(\alpha\wedge 1)\beta_j}{2(\alpha\wedge 1)\beta_j+1}}. \end{aligned} \tag{3.7.23}$$

Combining (3.7.18), (3.7.19), (3.7.20), (3.7.21), (3.7.22) and (3.7.23), 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 $\boldsymbol{\gamma}^\star \in \mathcal{F}_{[v_-,v_+]}(\alpha, \boldsymbol{\beta}, \mathbf{p}, L, \mathbf{R})$,

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

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

Proof of Corollary 3.4.2

We first present the following result which can be proved by a similar argument as the proof of Lemma 3.3.1.

Lemma 3.7.3. *Let* $\mathcal{C}_d = \left\{(c_1,\dots,c_d) \in \mathbb{R}^d,\ \sum_{j=1}^d |c_j| \leq 1\right\}$. *We denote* $\overline{\boldsymbol{S}}_{\mathcal{C}_d}$ *the collection of functions on* $[0,1]^d$ *of the form*

$$f(\boldsymbol{w}) = \frac{1}{2}\left(\langle c, \boldsymbol{w}\rangle + 1\right), \quad \textit{for all} \quad \boldsymbol{w} \in [0,1]^d, \tag{3.7.24}$$

with $c \in \mathcal{C}_d$ *and* $\boldsymbol{S}_{\mathcal{C}_d}$ *the collection of functions of the form in (3.7.24) 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 3.4.2. 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 $f \in \boldsymbol{S}_{(\boldsymbol{t},r)}^{\mathcal{H},l}$, $g : [0,1]^d \to [0,1]^l$ 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

$$\begin{aligned}
\|\boldsymbol{\gamma} \circ \boldsymbol{\gamma}' - (f \circ g) \vee v_-) \wedge v_+\|_\infty &\leq \|\boldsymbol{\gamma} \circ \boldsymbol{\gamma}' - f \circ g\|_\infty \\
&\leq \|\boldsymbol{\gamma} \circ \boldsymbol{\gamma}' - \boldsymbol{\gamma} \circ g\|_\infty + \|\boldsymbol{\gamma} \circ g - f \circ g\|_\infty \\
&\leq \left\| L \sum_{j=1}^{l} \left|\gamma_j' - g_j\right|^{\alpha_j \wedge 1} \right\|_\infty + \|\boldsymbol{\gamma} - f\|_\infty \\
&\leq L \sum_{j=1}^{l} \left\| \left|\gamma_j' - g_j\right|^{\alpha_j \wedge 1} \right\|_\infty + \|\boldsymbol{\gamma} - f\|_\infty. \qquad (3.7.25)
\end{aligned}$$

We apply Lemma 3.7.2 to (3.7.25) by taking $k = \infty$, μ the Lebesgue probability and $u(z) = z^{\alpha_j \wedge 1}$ for each $j \in \{1, \ldots, l\}$ and obtain

$$\|\boldsymbol{\gamma} \circ \boldsymbol{\gamma}' - (f \circ g \vee v_-) \wedge v_+\|_\infty \leq L \sum_{j=1}^{l} \left(\left\|\gamma_j' - g_j\right\|_\infty \right)^{\alpha_j \wedge 1} + \|\boldsymbol{\gamma} - f\|_\infty. \qquad (3.7.26)$$

We take $r = \max_{j=1,\ldots,l} \lfloor \alpha_j \rfloor \in \mathbb{N}$. By Proposition 3.7.4, for any $\boldsymbol{\gamma}^\star = \boldsymbol{\gamma} \circ \boldsymbol{\gamma}' \in \mathcal{G}_{[v_-,v_+]}(\boldsymbol{\alpha}, L)$ and all $\boldsymbol{t} = (t_1, \ldots, t_l) \in (\mathbb{N}^*)^l$, we have

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

where $C_{l,\boldsymbol{\alpha}}$ is a constant depending on l and $\boldsymbol{\alpha}$ only. Then by Lemma 3.4.1, $\boldsymbol{S}_{(\boldsymbol{t},r)}^{\mathcal{H},l}$ is dense in $\overline{\boldsymbol{S}}_{(\boldsymbol{t},r)}^{\mathcal{H},l}$ with respect to the supremum norm $\|\cdot\|_\infty$, we obtain

$$\inf_{f \in \boldsymbol{S}_{(\boldsymbol{t},r)}^{\mathcal{H},l}} \|\boldsymbol{\gamma} - f\|_\infty = \inf_{f \in \overline{\boldsymbol{S}}_{(\boldsymbol{t},r)}^{\mathcal{H},l}} \|\boldsymbol{\gamma} - f\|_\infty \leq C_{l,\boldsymbol{\alpha}} L \sum_{j=1}^{l} t_j^{-\alpha_j}. \qquad (3.7.27)$$

Therefore, following from (3.7.26), (3.7.27) and Lemma 3.7.3, for all $\boldsymbol{\alpha} \in (\mathbb{R}_+^*)^l$, $L > 0$, any $\boldsymbol{\gamma}^\star = \boldsymbol{\gamma} \circ \boldsymbol{\gamma}' \in \mathcal{G}_{[v_-,v_+]}(\boldsymbol{\alpha}, L)$ and all $\boldsymbol{t} = (t_1, \ldots, t_l) \in (\mathbb{N}^*)^l$,

$$\begin{aligned}
&\inf_{f \in \boldsymbol{S}_{(\boldsymbol{t},r)}^{\mathcal{H},l},\, g_j \in \boldsymbol{S}_{\mathcal{C}_d}} \|\boldsymbol{\gamma} \circ \boldsymbol{\gamma}' - (f \circ g \vee v_-) \wedge v_+\|_{2,P_W}^2 \\
\leq &\inf_{f \in \boldsymbol{S}_{(\boldsymbol{t},r)}^{\mathcal{H},l},\, g_j \in \boldsymbol{S}_{\mathcal{C}_d}} \|\boldsymbol{\gamma} \circ \boldsymbol{\gamma}' - (f \circ g \vee v_-) \wedge v_+\|_\infty^2 \\
\leq &C_{l,\boldsymbol{\alpha}} L^2 \left(\sum_{j=1}^{l} t_j^{-\alpha_j} \right)^2. \qquad (3.7.28)
\end{aligned}$$

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

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

and Proposition 3.4.2, we have

$$\begin{aligned} V^M_{(\boldsymbol{t},r)} + \Delta(\boldsymbol{t}, r) &\leq C_l \left[d + \left(\prod_{j=1}^{l} t_j \right) (r+1)^l \right] \log \left[\left(\sum_{j=1}^{l} t_j \right) + lr + l + 1 \right] \\ &\leq C_{l,\boldsymbol{\alpha}} \left[d + \left(\prod_{j=1}^{l} t_j \right) \right] \log \left[\left(\sum_{j=1}^{l} t_j \right) + lr + l + 1 \right], \end{aligned} \tag{3.7.29}$$

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 3.3.1, applying (3.2.8) together with the inequalities (3.7.28) and (3.7.29), 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,

$$\begin{aligned} &\mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}})\right] \\ \leq & C_\kappa \inf_{(\boldsymbol{t},r) \in (\mathbb{N}^*)^l \times \mathbb{N}} \left[\inf_{\widetilde{\boldsymbol{\gamma}} \in \boldsymbol{\Gamma}^M_{(\boldsymbol{t},r)}} \|\boldsymbol{\gamma} \circ \boldsymbol{\gamma}' - \widetilde{\boldsymbol{\gamma}}\|^2_{2,P_W} + \frac{\Delta(\boldsymbol{t}, r)}{n} + \frac{V^M_{(\boldsymbol{t},r)}}{n}(1 + \log n) \right] \\ \leq & C_{\kappa,l,\boldsymbol{\alpha}} (1 + \log n) \inf_{\boldsymbol{t} \in (\mathbb{N}^*)^l} \left[L^2 \left(\sum_{j=1}^{l} t_j^{-\alpha_j} \right)^2 + \frac{\prod_{j=1}^{l} t_j}{n} + \frac{d}{n} \right] \log(U), \end{aligned} \tag{3.7.30}$$

where $U = \sum_{j=1}^{l} t_j + lr + l + 1$. We then optimize the risk bound given on the right hand side of (3.7.30). For each $j \in \{1, \ldots, l\}$, 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 $\alpha_1, \ldots, \alpha_l$. Therefore, we have

$$L^2 \left(\sum_{j=1}^{l} t_j^{-\alpha_j} \right)^2 \leq l^2 L^2 (nL^2)^{-\frac{2\overline{\alpha}}{2\overline{\alpha}+l}} = l^2 L^{\frac{2l}{2\overline{\alpha}+l}} n^{-\frac{2\overline{\alpha}}{2\overline{\alpha}+l}}. \tag{3.7.31}$$

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

$$\frac{\prod_{j=1}^{l} t_j}{n} \leq \frac{1}{n} \tag{3.7.32}$$

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

$$\log(U) = \log \left[\left(\sum_{j=1}^{l} t_j \right) + lr + l + 1 \right] \leq C_{l,\boldsymbol{\alpha}}. \tag{3.7.33}$$

Otherwise,

$$\frac{\prod_{j=1}^{l} t_j}{n} \leq \frac{\prod_{j=1}^{l} 2(nL^2)^{\frac{\overline{\alpha}}{(2\overline{\alpha}+l)\alpha_j}}}{n} = \frac{2^l (nL^2)^{\frac{\overline{\alpha}}{2\overline{\alpha}+l}\sum_{j=1}^{l}\frac{1}{\alpha_j}}}{n}$$
$$\leq 2^l \frac{(nL^2)^{\frac{l}{2\overline{\alpha}+l}}}{n} \leq 2^l L^{\frac{2l}{2\overline{\alpha}+l}} n^{-\frac{2\overline{\alpha}}{2\overline{\alpha}+l}} \tag{3.7.34}$$

and

$$\log\left[\left(\sum_{j=1}^{l} t_j\right) + lr + l + 1\right] \leq \log\left[l\left(\prod_{j=1}^{l} t_j\right) + lr + l + 1\right]$$
$$\leq \log\left[C_l \left(nL^2\right)^{\frac{l}{2\overline{\alpha}+l}} + C_{l,\boldsymbol{\alpha}}\right]$$
$$\leq C_{l,\boldsymbol{\alpha}} \left(\log n \vee \log L^2 \vee 1\right). \tag{3.7.35}$$

Plugging (3.7.31), (3.7.32), (3.7.33), (3.7.34), (3.7.35) into (3.7.30), we have that whatever the distribution of W, for all $\boldsymbol{\alpha} \in (\mathbb{R}_+^*)^l$ and $L > 0$, any $\boldsymbol{\gamma}^\star \in \mathcal{G}_{[v_-,v_+]}(\boldsymbol{\alpha}, L)$

$$\mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}})\right] \leq C_{\kappa,l,\boldsymbol{\alpha}} \left(L^{\frac{2l}{2\overline{\alpha}+l}} n^{-\frac{2\overline{\alpha}}{2\overline{\alpha}+l}} + \frac{d}{n}\right) \left(\log n \vee \log L^2 \vee 1\right)^2, \tag{3.7.36}$$

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 $\mathcal{G}_{[v_-,v_+]}(\boldsymbol{\alpha}, L)$.

Proof of Corollary 3.5.1

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

$$\lim_{i\to+\infty} \|\boldsymbol{\gamma}_m^\star - \boldsymbol{\gamma}_i\|_\infty = \lim_{i\to+\infty} \left|\sum_{k=1}^{m} \left(t^k - t_i^k\right) \boldsymbol{\gamma}_1 \left(\boldsymbol{\gamma}_2^{\circ k}\right)(w)\right|$$
$$\leq K \lim_{i\to+\infty} \sum_{k=1}^{m} \left|t^k - t_i^k\right|$$
$$\leq \frac{Km(m+1)}{2} \lim_{i\to+\infty} |t - t_i| = 0, \tag{3.7.37}$$

since $\mathbb{Q}$ is dense in $\mathbb{R}$. Therefore, with the fact that $\boldsymbol{\gamma}^\star$ taking values in $[v_-, v_+]$ and (3.7.37), we have

$$\begin{aligned}\inf_{\overline{\boldsymbol{\gamma}}\in\boldsymbol{\Gamma}_{(l(m+1),p_1+p_2+2)}}\|\boldsymbol{\gamma}^\star-\overline{\boldsymbol{\gamma}}\|_\infty &\le \inf_{\overline{\boldsymbol{\gamma}}\in\boldsymbol{S}_{(l(m+1),p_1+p_2+2)}}\|\boldsymbol{\gamma}^\star-\overline{\boldsymbol{\gamma}}\|_\infty\\ &\le \|\boldsymbol{\gamma}^\star-\boldsymbol{\gamma}^\star_m\|_\infty+\inf_{\overline{\boldsymbol{\gamma}}\in\boldsymbol{S}_{(l(m+1),p_1+p_2+2)}}\|\boldsymbol{\gamma}^\star_m-\overline{\boldsymbol{\gamma}}\|_\infty\\ &\le \sup_{w\in[0,1]}\left|\sum_{k=m+1}^{+\infty}t^k\boldsymbol{\gamma}_1\left(\boldsymbol{\gamma}_2^{\circ k}(w)\right)\right|\\ &\le C_{t,K}|t|^{m+1}, \end{aligned}\tag{3.7.38}$$

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 $\boldsymbol{S}_{(l(m+1),p_1+p_2+2)}$ is a subset of $\overline{\boldsymbol{S}}_{(l(m+1),p_1+p_2+2)}$, Proposition 3.5.1 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

$$V_{(l(m+1),p_1+p_2+2)}\le Cl_0\left[p_0^2(l_0-1)+p_0(l_0+2)+1\right]\log\left[(l_0+1)\left(\frac{p_0l_0}{2}+1\right)\right],$$

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

$$V_{(l(m+1),p_1+p_2+2)}+\Delta(l(m+1),p_1+p_2+2)\le C_{l,\mathbf{p}}(m+1)^3,\tag{3.7.39}$$

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 3.3.1, applying (3.2.8) together with the inequalities (3.7.38) and (3.7.39), 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 $K\ge 0$, for any $\boldsymbol{\gamma}^\star\in\mathcal{F}_{[v_-,v_+]}(t,l,\mathbf{p},K)$,

$$\begin{aligned}\mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star},R_{\widehat{\boldsymbol{\gamma}}})\right]\le& C_{\kappa,\mathbf{p},l}\inf_{m\in\mathbb{N}^*}\left[\inf_{\overline{\boldsymbol{\gamma}}\in\boldsymbol{\Gamma}_{(l(m+1),p_1+p_2+2)}}\|\boldsymbol{\gamma}^\star-\overline{\boldsymbol{\gamma}}\|^2_{2,P_W}+\frac{(m+1)^3}{n}(1+\log n)\right]\\ \le& C_{\kappa,\mathbf{p},l,t,K}\inf_{m\in\mathbb{N}^*}\left[|t|^{2(m+1)}+\frac{(m+1)^3}{n}(1+\log n)\right].\end{aligned}\tag{3.7.40}$$

Now we only need to optimize the right hand side of (3.7.40). If $|t|^4\le 1/n$, we choose $m=1$ so that

$$\mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star},R_{\widehat{\boldsymbol{\gamma}}})\right]\le C_{\kappa,\mathbf{p},l,t,K}\frac{1}{n}(1+\log n).\tag{3.7.41}$$

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

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

With this choice, $|t|^{2(m+1)} \leq 1/n$ and we derive from (3.7.40),

$$\mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}})\right] \leq C_{\kappa,\mathbf{p},l,t,K} \frac{(1+\log n)^4}{n}. \tag{3.7.42}$$

Combining the results in (3.7.41) and (3.7.42), 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

$$\mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}})\right] \leq C_{\kappa,\mathbf{p},l,t,K} \frac{1}{n}(1+\log n)^4. \tag{3.7.43}$$

Then the conclusion follows by taking the supremum over $\mathcal{F}_{[v_-,v_+]}(t,l,\mathbf{p},K)$ on both sides of (3.7.43).

Proof of Corollary 3.5.2

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 3.7.4. *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 $l \in \{1,\dots,L\}$

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

and f_l^-, $l \in \{1,\dots,L+1\}$, 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 $\boldsymbol{w} \in [0,1]^d$

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

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 $j \in \{0, \dots, k\}$, $\boldsymbol{\alpha} \in (\mathbb{R}_+^*)^{k+1}$ 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

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

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 3.7.4, we have

$$\begin{aligned}
\inf_{\overline{\boldsymbol{\gamma}} \in \overline{\boldsymbol{\Gamma}}_{(L,p,\boldsymbol{s})}} \|\boldsymbol{\gamma}^\star - \overline{\boldsymbol{\gamma}}\|_\infty^2 \leq \inf_{\overline{\boldsymbol{\gamma}} \in \overline{\boldsymbol{S}}_{(L,p,\boldsymbol{s})}} \|\boldsymbol{\gamma}^\star - \overline{\boldsymbol{\gamma}}\|_\infty^2 &\leq \inf_{\overline{\boldsymbol{\gamma}} \in \overline{\boldsymbol{S}}'_{(L,p,\boldsymbol{s})}} \|\boldsymbol{\gamma}^\star - \overline{\boldsymbol{\gamma}}\|_\infty^2 \\
&\leq \inf_{\overline{\boldsymbol{\gamma}} \in \overline{\boldsymbol{S}}'_{(L,p,\boldsymbol{s})}} \|\boldsymbol{\gamma}^\star - \overline{\boldsymbol{\gamma}}\|_\infty^2 \\
&\leq C_{k,\mathbf{d},\mathbf{t},\boldsymbol{\alpha},K} \max_{i=0,\dots,k} n^{-\frac{2\alpha'_i}{2\alpha'_i + t_i}}. \qquad (3.7.44)
\end{aligned}$$

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 3.5.1 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

$$\begin{aligned}
V_{(L,p,\boldsymbol{s})} &\leq C_{k,\mathbf{d},\mathbf{t},\boldsymbol{\alpha}} n\phi_n (\log n)^2 \log\left[(L+1)\left(\frac{pL}{2}+1\right)\right] \\
&\leq C_{k,\boldsymbol{d},\boldsymbol{t},\boldsymbol{\alpha}} n\phi_n (\log n)^3. \qquad (3.7.45)
\end{aligned}$$

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

$$\begin{aligned}
\Delta(L,p,\boldsymbol{s}) &\leq \|\boldsymbol{s}\|_0 \log(2e\overline{p}) + p + L \\
&\leq C_{k,\boldsymbol{d},\boldsymbol{t},\boldsymbol{\alpha}} (n\phi_n \log n \log \overline{p} + n\phi_n + \log_2 n) \\
&\leq C_{k,\boldsymbol{d},\boldsymbol{t},\boldsymbol{\alpha}} n\phi_n (\log n)^2. \qquad (3.7.46)
\end{aligned}$$

Under Assumption 3.3.1, applying (3.2.8) together with (3.7.44), (3.7.45) and (3.7.46), whatever the distribution of W, we derive that for all $k \in \mathbb{N}^*$, $K \geq 0$, $\mathbf{d} \in (\mathbb{N}^*)^{k+1}$, $\mathbf{t} \in$

$(\mathbb{N}^*)^{k+1}$ with $t_j \le d_j$ 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

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

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

Proof of Corollary 3.6.1

We note that according to Proposition 1.5.1, the collection of models $\{\overline{\boldsymbol{\Gamma}}_m,\ m \in \mathcal{M}\}$ satisfies Assumption 3.2.1 with $V_m = |m| + 1$. By Lemma 3.6.1, the associated weights $\Delta(m)$ satisfy inequality (3.2.2) with $\Sigma \le 1 + \pi^2/6$. Moreover, for each $m \in \mathcal{M}$, the countable subset $\boldsymbol{\Gamma}_m$ is dense in $\overline{\boldsymbol{\Gamma}}_m$ for the topology of pointwise convergence so that $h(Q^\star, \mathscr{Q}_m) = h(Q^\star, \overline{\mathscr{Q}}_m)$.

We apply (3.2.8) and derive that whatever the distribution of W, the resulted estimator $R_{\widehat{\boldsymbol{\gamma}}}$ satisfies

$$\mathbb{E}\left[h^2(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}})\right] \le c_2(c_3 + \Sigma)(\mathcal{B}_o \wedge \mathcal{B}_c), \tag{3.7.47}$$

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

$$\begin{aligned}
\mathcal{B}_o &\le \inf_{m \in \mathcal{M}_o}\left[h^2(R_{\boldsymbol{\gamma}^\star}, \overline{\mathscr{Q}}_m) + \frac{|m|+1}{n} + \frac{|m|+1}{n}\left[1 + \log_+\left(\frac{n}{|m|+1}\right)\right]\right] \\
&\le 2\inf_{m \in \mathcal{M}_o}\left[h^2(R_{\boldsymbol{\gamma}^\star}, \overline{\mathscr{Q}}_m) + \frac{|m|+1}{n}\left[1 + \log_+\left(\frac{n}{|m|+1}\right)\right]\right] = 2\mathscr{B}_o.
\end{aligned} \tag{3.7.48}$$

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

$$\begin{aligned}
\mathcal{B}_c &\le \inf_{m \in \mathcal{M}\setminus\mathcal{M}_o}\left[h^2(R_{\boldsymbol{\gamma}^\star}, \overline{\mathscr{Q}}_m) + \frac{|m|+1}{n}\left[1 + \log\left(\frac{2ep}{|m|+1}\right) + \log_+\left(\frac{n}{|m|+1}\right)\right]\right] \\
&\le 2\inf_{m \in \mathcal{M}\setminus\mathcal{M}_o}\left[h^2(R_{\boldsymbol{\gamma}^\star}, \overline{\mathscr{Q}}_m) + \frac{|m|+1}{n}\left[1 + \log\left(\frac{(2p)\vee n}{|m|+1}\right)\right]\right].
\end{aligned} \tag{3.7.49}$$

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

$$\log_+\left(\frac{n}{|m|+1}\right) \leq \log\left(\frac{(2p)\vee n}{|m|+1}\right). \tag{3.7.50}$$

Combining (3.7.47), (3.7.48), (3.7.49) and (3.7.50), 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.

3.7.2 Proofs of lemmas

Proof of Lemma 3.3.1

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

$$\widetilde{\boldsymbol{\gamma}}(\boldsymbol{w}) = \sum_{(r_1,\ldots,r_d)\in\{0,\ldots,r\}^d} \widetilde{\gamma}_{(r_1,\ldots,r_d)} \prod_{j=1}^d {w_j}^{r_j}, \tag{3.7.51}$$

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

$$\widetilde{\boldsymbol{\gamma}}_l(\boldsymbol{w}) = \sum_{(r_1,\ldots,r_d)\in\{0,\ldots,r\}^d} \widetilde{\gamma}^l_{(r_1,\ldots,r_d)} \prod_{j=1}^d {w_j}^{r_j}, \tag{3.7.52}$$

with $\widetilde{\gamma}^l_{(r_1,\ldots,r_d)} \in \mathbb{Q}$, for all $0 \leq r_j \leq r$, $1 \leq j \leq d$ 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 $l \geq N_\epsilon$,

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

Hence, for any $\widetilde{\gamma}$ defined by (3.7.51) and all $\epsilon > 0$, there exists a $N_\epsilon > 0$ and a sequence of functions $\{\widetilde{\gamma}_l\}_{l\in\mathbb{N}}$ defined by (3.7.52) such that for all $l \geq N_\epsilon$,

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

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 $g \in \overline{\boldsymbol{S}}^{\mathcal{B},d}_{(\boldsymbol{s},r)}$

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

□

Proof of Lemma 3.3.2

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 $\cup_{\boldsymbol{s}\in\mathbb{N}^d} M^{\mathcal{B},d}_{\boldsymbol{s}} \subset \cup_{D\in\mathbb{N}^*} M^d_D$, we have

$$\sum_{(\boldsymbol{s},r)\in\mathcal{M}} \exp\left[-\log(8d)\prod_{j=1}^d 2^{s_j} - r\right] \leq \sum_{r\in\mathbb{N}}\sum_{D\in\mathbb{N}^*}\sum_{\pi\in M^d_D} e^{-\log(8d)|\pi|-r}, \tag{3.7.53}$$

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 (3.7.53) that

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

□

Proof of Lemma 3.4.2

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 $M^1_{D_j}$ represent the set of partitions which divide $[0,1]$ into D_j subintervals. Recall that $M^{\mathcal{B},1}_s$ denotes the dyadic partition of $[0,1]$ into 2^s subintervals, hence we have for any $j\in\{1,\ldots,d\}$, $\cup_{s_j\in\mathbb{N}} M^{\mathcal{B},1}_{s_j} \subset \cup_{D_j\in\mathbb{N}^*} M^1_{D_j}$. As an immediate consequence,

$$\begin{aligned}
&\sum_{(\boldsymbol{s},t,r)\in\mathbb{N}^d\times\mathbb{N}^*\times\mathbb{N}} e^{-\Delta(\boldsymbol{s},t,r)} \\
&= \sum_{(\boldsymbol{s},t,r)\in\mathbb{N}^d\times\mathbb{N}^*\times\mathbb{N}} \exp\left[-t - \sum_{j=1}^{d} 2^{s_j}\log 8 - r\right] \\
&\leq \sum_{r\in\mathbb{N}} e^{-r}\left[\prod_{j=1}^{d}\left(\sum_{D_j\in\mathbb{N}^*}\sum_{\pi_j\in M^1_{D_j}} e^{-|\pi_j|\log 8}\right)\right]\left(\sum_{t\in\mathbb{N}^*} e^{-t}\right),
\end{aligned} \tag{3.7.54}$$

where $|\pi_j|$ denotes the cardinality of segments given by the partition π_j. Moreover, as we have mentioned in the proof of Lemma 3.3.2, it follows from Proposition 5 in Akakpo (2012) that $|M^1_{D_j}| \leq 4^{D_j}$ for $j\in\{1,\ldots,d\}$. Therefore, we derive from (3.7.54) that

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

□

Proof of Lemma 3.4.3

Proof.

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

□

Proof of Lemma 3.5.1

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

$$\sum_{k=0}^{|m|}\binom{p}{k} \leq \left(\frac{ep}{|m|}\right)^{|m|}. \tag{3.7.55}$$

Proof. First, we note that

$$\begin{aligned}&\sum_{(L,p,\boldsymbol{s})\in(\mathbb{N}^*)^2\times\{0,1\}^{\overline{p}}} e^{-\Delta(L,p,\boldsymbol{s})} \\ &= \sum_{L\in\mathbb{N}^*} e^{-L}\left[\sum_{p\in\mathbb{N}^*} e^{-p}\left(1 + \sum_{\boldsymbol{s}\in\{0,1\}^{\overline{p}}\setminus\{0\}^{\overline{p}}} \exp\left[-\|\boldsymbol{s}\|_0 \log\left(\frac{2e\overline{p}}{\|\boldsymbol{s}\|_0}\right)\right]\right)\right] \\ &\leq \sum_{L\in\mathbb{N}^*} e^{-L}\left\{\sum_{p\in\mathbb{N}^*} e^{-p}\left[1 + \sum_{s=1}^{\overline{p}}\binom{\overline{p}}{s}\exp\left(-s\log\left(\frac{2e\overline{p}}{s}\right)\right)\right]\right\}.\end{aligned} \tag{3.7.56}$$

By (3.7.55), we know for any $1 \leq s \leq \overline{p}$,

$$\binom{\overline{p}}{s} \leq \sum_{h=0}^{s}\binom{\overline{p}}{h} \leq \left(\frac{e\overline{p}}{s}\right)^s. \tag{3.7.57}$$

Plugging (3.7.57) into (3.7.56), we obtain

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

□

Proof of Lemma 3.6.1

Proof. By (3.7.55), we derive that

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

□

3.7.3 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 3.7.5. *Suppose $f_1(\cdot), f_2(\cdot), \ldots, f_T(\cdot)$ are fixed polynomials of degree at most d in $s \leq T$ variables. Define*

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

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

The second Lemma is the weighted AM-GM Inequality.

Lemma 3.7.6 (Weighted AM-GM Inequality)**.** *If $0 \leq c_i \in \mathbb{R}$ and $0 \leq \lambda_i \in \mathbb{R}$ for all $i = 1, \ldots, K$ such that $\sum_{i=1}^K \lambda_i = 1$, then*

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

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

Lemma 3.7.7. *Suppose that $2^m \leq 2^t (mr/w)^w$ for some $r \geq 16$ and $m \geq w \geq t \geq 0$. Then, $m \leq t + w \log_2(2r \log_2 r)$.*

Proof of Proposition 3.4.1

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 3.3 and $\overline{\boldsymbol{S}}^{\mathcal{H},1}_{(t,r)}$ in Section 3.4. 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 3.2.

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.

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

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 3.4 is the regular partition of $[0,1]$ into t subintervals. Let $\{b_1, \ldots, b_{t-1}\}$ 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 3.7.5 to the collection of polynomials

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

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 $\mathcal{S}_1 = \{S_1, \ldots, S_{N_1}\}$ 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 $i \in \{1, \ldots, m\}$, 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 3.7.5 again, on each $S \in \mathcal{S}_1$,

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

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 $\mathcal{S}_2 = \{S_1, \ldots, S_{N_1 N_2}\}$ over $\mathbb{R}^s$ with at most $N_1 N_2$ pieces such that within each region $S \in \mathcal{S}_2$,

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

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 $(\boldsymbol{w}_1, v_1), \ldots, (\boldsymbol{w}_m, v_m) \in \mathscr{W} \times \mathbb{R}$, we have

$$N(m) \leq \sum_{k=1}^{N_1 N_2} |\{(\operatorname{sgn}(h_{\boldsymbol{w}_1}(a) - v_1), \ldots, \operatorname{sgn}(h_{\boldsymbol{w}_m}(a) - v_m)),\ a \in S_k\}| \\ \leq N_1 N_2 \leq 4 \left(c_1^{\lambda_1} c_2^{\lambda_2}\right)^{t(r+1)+2\sum_{j=1}^d |\pi_j|(r+1)}. \tag{3.7.58}$$

Applying Lemma 3.7.6 to (3.7.58), we derive that

$$N(m) \leq N_1 N_2 \leq 4 \left(\lambda_1 c_1 + \lambda_2 c_2\right)^{t(r+1)+2\sum_{j=1}^d |\pi_j|(r+1)} \\ \leq 4 \left(\frac{2em(t+r+2)}{t(r+1) + 2\sum_{j=1}^d |\pi_j|(r+1)} \right)^{t(r+1)+2\sum_{j=1}^d |\pi_j|(r+1)}. \tag{3.7.59}$$

From the definition of VC-dimension together with (3.7.59),

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

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 3.7.7 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 $V^{A}_{(\boldsymbol{\pi},t,r)} \leq V^{\widetilde{A}}_{(\boldsymbol{\pi},t,r)}$. □

Proof of Proposition 3.4.2

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

$$\boldsymbol{\gamma}(\boldsymbol{w}) = f\left(g_1(\boldsymbol{w}), \ldots, g_l(\boldsymbol{w})\right), \text{ for all } \boldsymbol{w} \in [0,1]^d \tag{3.7.60}$$

where $f \in \overline{\boldsymbol{S}}^{\mathcal{H},l}_{(\boldsymbol{t},r)}$, $g_j(\boldsymbol{w}) = [((\langle a_j, \boldsymbol{w}\rangle + 1)/2) \vee 0] \wedge 1$ 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 3.4 and (3.7.60), 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 $\widetilde{\mathbf{\Gamma}}^M_{(\boldsymbol{t},r)}$ induced by the parameters vector $a \in \mathbb{R}^s$ hence we can rewrite $\widetilde{\mathbf{\Gamma}}^M_{(\boldsymbol{t},r)} = \{f_a,\ a \in \mathbb{R}^s\}$. 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.

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

The idea is similar to the proof of Proposition 3.4.1 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 3.7.5.

We initialise our partition of $\mathbb{R}^s$ with $\mathcal{S}_0 = \{\mathbb{R}^s\}$. We recall that for a given $\boldsymbol{t} = (t_1, \ldots, t_l) \in (\mathbb{N}^*)^l$, $M^{\mathcal{H},l}_{\boldsymbol{t}}$ defined in Section 3.4 is the partition of $[0,1]^l$, where in all the directions $j \in \{1, \ldots, l\}$, the interval $[0,1]$ is divided into t_j regular subintervals. Let $\{b^j_1, \ldots, b^j_{t_j - 1}\}$ be the breakpoints on the interval $(0,1)$ in the j-th direction given by the partition $M^{\mathcal{H},l}_{\boldsymbol{t}}$ and denote $b^j_0 = 0$, $b^j_{t_j} = 1$ for all $j \in \{1, \ldots, l\}$. We consider the collection of polynomials

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

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 3.7.5. Therefore, we partition $\mathbb{R}^s$ into N_1 pieces with the refined partition $\mathcal{S}_1 = \{S_1, \ldots, S_{N_1}\}$ 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 $i \in \{1, \ldots, m\}$, $h_{\boldsymbol{w}_i}(a)$ 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 3.7.5 again, on each $S \in \mathcal{S}_1$,

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

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 $(\boldsymbol{w}_1, v_1), \dots, (\boldsymbol{w}_m, v_m) \in \mathscr{W} \times \mathbb{R}$, we have

$$\begin{aligned} N(m) &\leq \sum_{k=1}^{N_1 N_2} \left| \left\{ (\operatorname{sgn}(h_{\boldsymbol{w}_1}(a) - v_1), \dots, \operatorname{sgn}(h_{\boldsymbol{w}_m}(a) - v_m)),\ a \in S_k \right\} \right| \\ &\leq N_1 N_2 \leq 4 \left(c_1^{\lambda_1} c_2^{\lambda_2} \right)^{2ld + \left(\prod_{j=1}^l t_j\right)(r+1)^l}. \end{aligned} \tag{3.7.61}$$

Applying Lemma 3.7.6 to (3.7.61), we derive that

$$\begin{aligned} N(m) &\leq N_1 N_2 \leq 4 \left(\lambda_1 c_1 + \lambda_2 c_2\right)^{2ld + \left(\prod_{j=1}^l t_j\right)(r+1)^l} \\ &\leq 4 \left[\frac{2em\left(\sum_{j=1}^l t_j + lr + l + 1\right)}{2ld + \left(\prod_{j=1}^l t_j\right)(r+1)^l} \right]^{2ld + \left(\prod_{j=1}^l t_j\right)(r+1)^l}. \end{aligned} \tag{3.7.62}$$

From the definition of VC-dimension together with (3.7.62),

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

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

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

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

Proof of Proposition 3.5.1

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, $l \in \{0,\dots,L\}$. For each $l \in \{0,\dots,L\}$, 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 $(\boldsymbol{w}_1, t_1), \dots, (\boldsymbol{w}_m, t_m) \in \mathscr{W} \times \mathbb{R}$, 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.

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

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 $h_{\boldsymbol{w}_j}(a) - t_j$ $j \in \{1,\dots,m\}$ are all fixed polynomials of a with a bounded degree.

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

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

$$\begin{cases} |\mathcal{S}_l| = |\mathcal{S}_{l-1}| & , \quad \text{if } \sum_{i=0}^{l-1} s_i = 0, \\ |\mathcal{S}_l| \leq 2\left(\dfrac{2emlp}{\sum_{i=0}^{l-1} s_i}\right)^{\sum_{i=0}^{l-1} s_i} |\mathcal{S}_{l-1}| & , \quad \text{if } \sum_{i=0}^{l-1} s_i \neq 0. \end{cases} \tag{3.7.63}$$

2. For each $l \in \{1,\dots,L\}$ and each $S \in \mathcal{S}_{l-1}$, when a varies in S, the input to each node in response to each $\boldsymbol{w}_j$, $j \in \{1,\dots,m\}$ 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 $\boldsymbol{w}_j$, $j \in \{1,\dots,m\}$, 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 3.7.5 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 (3.7.63), we derive

$$|\mathcal{S}_L| \leq \prod_{l=1}^{L} N_l. \tag{3.7.64}$$

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 3.7.5 again, we have for any $S \in \mathcal{S}_L$,

$$\begin{aligned} N_{L+1} &= \left|\{(\operatorname{sgn}(h_{\boldsymbol{w}_1}(a) - t_1), \ldots, \operatorname{sgn}(h_{\boldsymbol{w}_m}(a) - t_m)),\ a \in S\}\right| \\ &\leq 2\left(\frac{2em(L+1)}{s}\right)^{s}. \end{aligned} \tag{3.7.65}$$

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 (3.7.64) and (3.7.65). 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 $(\boldsymbol{w}_1, t_1), \ldots, (\boldsymbol{w}_m, t_m) \in \mathscr{W} \times \mathbb{R}$,

$$\begin{aligned} N(m) &\leq \sum_{k=1}^{N} \left| \{(\operatorname{sgn}(h_{\boldsymbol{w}_1}(a) - t_1), \ldots, \operatorname{sgn}(h_{\boldsymbol{w}_m}(a) - t_m)),\ a \in S_k\} \right| \\ &\leq \prod_{l=1}^{L+1} N_l = 2^{L+2-\bar{l}} \left[\prod_{l=\bar{l}}^{L+1} \left(\frac{2emlp_l}{\sum_{i=0}^{l-1} s_i} \right)^{\sum_{i=0}^{l-1} s_i} \right]. \end{aligned} \tag{3.7.66}$$

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 3.7.6 to (3.7.66) 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 3.7.7 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. □

Chapter 4

Estimation by estimator selection with application to changepoint detection

4.1 Introduction

In this chapter, we study the same estimation problem introduced in Chapter 1 with another estimation strategy based on estimator selection. Let us briefly recall our statistical setting described in Chapter 1: we observe n pairs of independent (but not necessarily i.i.d.) random variables, i.e. $X_i = (W_i, Y_i)$ for $i \in \{1, \dots, n\}$, with values in a measurable product space $(\mathscr{W} \times \mathscr{Y}, \mathcal{W} \otimes \mathcal{Y})$. For each i, we assume the conditional distribution $Q_i^\star(w_i)$ of Y_i exists and is given by the value of a measurable function $Q_i^\star \in \mathscr{Q}_{\mathscr{W}}$ at the point w_i, where $\mathscr{Q}_{\mathscr{W}}$ denotes the set of all measurable mappings from $(\mathscr{W}, \mathcal{W})$ into $(\mathscr{T}, \mathcal{T})$ (see Section 2.2). Our statistical interest lies in estimating $\mathbf{Q}^\star$ the n conditional distributions $Q_i^*(w_i)$ of Y_i given $W_i = w_i$ on the basis of the observations $\boldsymbol{X} = (X_1, \dots, X_n)$. We do as if there exists an unknown function $\boldsymbol{\gamma}^\star$ on $\mathscr{W}$ such that for each $i \in \{1, ..., n\}$, the conditional distribution of Y_i given $W_i = w_i$ belongs to a one-parameter exponential family with parameter $\boldsymbol{\gamma}^\star(w_i) \in \mathbb{R}$. Throughout this chapter, unless otherwise specified, we assume the exponential family $\overline{\mathscr{Q}}$ has been parametrized under its general form and write it as $\overline{\mathscr{Q}} = \{R_\gamma = r_\gamma \cdot \nu,\ \gamma \in J\}$ with the densities r_γ given in (2.3.12).

In Chapter 2, we have introduced an estimation procedure based on one model, where the idea comes from ρ-estimation (Baraud et al. (2017) and Baraud and Birgé (2018)). As we have commented in Chapter 3, this approach performs well if one knows a suitable model for the potential regression function $\boldsymbol{\gamma}^\star$ in advance which means a model can provide a good enough approximation and is also not too complicated. But such a model can be difficult to design in some situations where only few prior information is known. A safer

strategy has been discussed in Chapter 3 where the problem is solved by a model selection procedure. However, one disadvantage of this strategy is the expensive numerical cost especially when the number of the models becomes large. At this point, an interesting problem could be can we come up with a new estimation strategy overcoming the above two limitations? This is to say the desired strategy should be capable of comparing estimators from several different models with a reasonable numerical cost.

Another motivation comes from the context of changepoint detection problem in exponential families. To be more precise, statisticians consider a specific situation in our setting where $W_i = (i-1)/n$ are deterministic and $\boldsymbol{\gamma}^\star : [0,1) \to J \subset \mathbb{R}$ is a right-continuous step function with an unknown number $N-1$ of changepoints (i.e. N segments, $N \geq 1$). Frick et al. (2013) proposed a simultaneous multiscale changepoint estimator (SMUCE for short). For each candidate estimator, Frick et al. (2013) designed a multiscale statistic to evaluate the maximum over the local likelihood ratio statistics on all discrete intervals such that the estimator is constant on these intervals with some value. Then provided a threshold q, the quantity N is estimated by $\widehat{N}(q)$ which is the number of segments of the estimators satisfying their threshold condition with the minimal segments. Finally, their estimator is the likelihood maximizer over a constrained set in which all the estimators satisfy the threshold condition with exact $\widehat{N}(q)$ segments. Cleynen and Lebarbier (2014, 2017) considered partitions given by the pruned dynamic programming algorithm (Rigaill (2015)) and proposed a penalized log-likelihood estimator following the work of constructing the penalty function done by L. Birgé and P. Massart (see Barron et al. (1999) and Birgé and Massart (1997) for instance). They also showed that the resulting estimator satisfies some oracle inequalities. Similar to the existing literature we mentioned in Chapter 1, these two methods are also both, more or less, based on the maximum likelihood estimation. When there are outliers presenting in the observations, both of the two procedures infer extra changepoints to fit the outliers while identifying the true ones in the signal. For this point, we shall illustrate it in a more straightforward way in the simulation part of this chapter. A natural question is can we find a procedure to enhance the stability of their estimators?

Besides these procedures specially designed for the changepoint detection problem in exponential families, detecting changes in the characteristics of a sequence of observed random variables has a long history and experienced a renaissance in recent years boosted by a flourishing development in bioinformatics (e.g. Olshen et al. (2004), Huang et al. (2005), Tibshirani and Wang (2007), Zhang and Siegmund (2007) and Muggeo and Adelfio (2010)). It also has attracted attention from other fields including climatology (e.g. Reeves et al. (2007) and Gallagher et al. (2013)), financial econometrics (e.g. Spokoiny (2009)) and signal processing (e.g. Blythe et al. (2012) and Hotz et al. (2013)), among many others. Within the regime of univariate mean changepoint detection, theoretical analysis has been established recently by Verzelen et al. (2020) and Wang et al. (2020). A recently selective

review of the related literature can be found in Truong et al. (2020). We only mention some representative procedures here. Scott and Knott (1974) proposed a binary segmentation (BS for short) method to detect the changes in means. A modified procedure circular binary segmentation (CBS for short) was provided by Olshen et al. (2004) and then a faster algorithm was given in Venkatraman and Olshen (2007) which has achieved a big success in genome analysis. Later, to enhance the robustness to departures from standard model assumptions, another method (denoted as cumSeg in the sequel) had been tailor-made by Muggeo and Adelfio (2010) to detect changes in genomic sequences. To reduce the complexity for computation, the pruned exact linear time method was proposed (PELT for short) by Killick et al. (2012) where they also showed PELT leads to a substantially more accurate result than BS. Wild binary segmentation (WBS for short) is an approach proposed by Fryźlewicz (2014) based on a development of BS and it becomes quite popular nowadays due to its nice performance and easy implementation. Aimed at improving SMUCE (Frick et al. (2013)) especially under the situation with low signal-to-noise ratio or with many changepoints compared to the length of the observations, Li et al. (2016) proposed an alternative multiscale segmentation method (denoted as FDR in the sequel) by controlling the false discovery rate of the whole segmentation. In the direction of being robust in the presence of outliers, Fearnhead and Rigaill (2019) proposed an algorithm (denoted as robseg in the sequel) based on the idea of adapting existing penalized cost methods to some loss functions which are less sensitive to the outliers. Two examples of the loss functions to which their procedure applies are Huber loss and biweight loss. In practice, based on the same observations, different approaches mentioned above may give different estimators. As it was point out by the comparison study in Fearnhead and Rigaill (2020), it is rather rare that one particular method uniformly outperforms another. Given so many experts' suggestions, a realistic and also interesting question is which one we should pick? Or in another word, can we let the data decide the preference of several (possibly random) estimators case by case?

These three problems mentioned above are the main motivations to propose the content in this chapter. In fact, we shall see that all of them can be solved simultaneously by an estimation strategy based on a data-driven estimator selection (denoted as ES in the sequel). More precisely, given the observations $\boldsymbol{X} = (X_1, \ldots, X_n)$, we assume to have at disposal an arbitrary but at most countable collection of piecewise constant (possibly random) candidates for the potential regression function $\boldsymbol{\gamma}^\star$ mapping $\mathscr{W}$ into J written as $\widehat{\boldsymbol{\Gamma}} = \{\widehat{\boldsymbol{\gamma}}_\lambda(\boldsymbol{X}),\ \lambda \in \Lambda\}$. The dependency of each candidate in $\widehat{\boldsymbol{\Gamma}}$ on the observations $\boldsymbol{X}$ can be unknown. We design an algorithm to compare these candidates in $\widehat{\boldsymbol{\Gamma}}$ pair by pair based on the same observations $\boldsymbol{X}$ and let the data choose the desired one denoted as $\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}(\boldsymbol{X})$ (or $\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}$ for short). Once obtaining $\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}$, our estimator of $\mathbf{Q}^\star$ is given by $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}} = (R_{\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}}, \ldots, R_{\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}}) \in \boldsymbol{\mathscr{Q}}_{\mathscr{W}} = \mathscr{Q}_{\mathscr{W}}^n$. With a slight abuse of language, sometimes we also call $\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}$ an estimator of $\boldsymbol{\gamma}^\star$ though we know that such a $\boldsymbol{\gamma}^\star$ may not necessarily exist. It is also

worthy to emphasis hereby that besides the independence, we assume nothing about the distributions of the covariates W_i which therefore can be unknown.

To evaluate the performance of the selected estimator $\mathbf{R}_{\widehat{\gamma}_{\widehat{\lambda}}}$, we need to introduce a loss function and we use the same Hellinger-type (pseudo) distance $\mathbf{h}$ on $\mathscr{Q}_{\mathscr{W}}$ defined by (2.2.2) as we do in Chapter 2 and 3. In particular, in the context of changepoint detection problem in exponential families where W_i are deterministic, the loss function $\mathbf{h}$ is nothing but the sum (from $i = 1$ to n) of the Hellinger distance between each two probabilities. From this point of view, unlike the typical methods detecting changes for some parameter of a distribution (for example detecting changes in means for Gaussian and Poisson distributions), our approach validates the changes along the sequence if there are abrupt variations with respect to the distribution.

The remainder of this chapter is organized as follows. We present our estimator selection procedure as well as the theoretical properties of the resulting estimator in Section 4.2. In Section 4.3, we explain how to apply this procedure to changepoint detection problem in exponential families. Section 4.4 is devoted to a comparative simulation study for illustrating the practical performance of the selected estimator. The performance on two real datasets (DNA copy numbers and British coal disasters) is exhibited in Section 4.5. Finally, all the proofs in this chapter are left to Section 4.6 and information about the testing signals we used in Section 4.4 is provided in Section 4.7.

4.2 Estimator selection strategy

As already mentioned, given the observations $\boldsymbol{X} = (X_1, \ldots, X_n)$, we assume that we have at disposal an at most countable (possibly random) candidates $\widehat{\mathbf{\Gamma}} = \{\widehat{\boldsymbol{\gamma}}_\lambda(\boldsymbol{X}),\ \lambda \in \Lambda\}$ for $\boldsymbol{\gamma}^\star$, where for each $\lambda \in \Lambda$, $\widehat{\boldsymbol{\gamma}}_\lambda$ is piecewise constant on $\mathscr{W}$. This $\widehat{\mathbf{\Gamma}}$ may contain the estimators based on the minimization of some criterions, estimators based on Bayes procedures or just simple guesses by some experts. The dependency of these estimators with respect to the observations $\boldsymbol{X}$ can be unknown. Our goal is to select some $\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}(\boldsymbol{X})$ among the family $\widehat{\mathbf{\Gamma}} = \{\widehat{\boldsymbol{\gamma}}_\lambda(\boldsymbol{X}),\ \lambda \in \Lambda\}$ based on the same observations $\boldsymbol{X}$ such that the risk of our estimator is as close as possible to the quantity $\inf_{\lambda\in\Lambda} \mathbb{E}\left[\mathbf{h}^2(\mathbf{Q}^\star, \mathbf{R}_{\widehat{\gamma}_\lambda})\right]$.

4.2.1 Estimator selection procedure

Let $\mathcal{M}$ be a finite or countable set of partitions on $\mathscr{W}$. We begin with a family of collections $\{\mathbf{\Gamma}_m,\ m \in \mathcal{M}\}$ indexed by the partition m on $\mathscr{W}$, where for each $m \in \mathcal{M}$, $\mathbf{\Gamma}_m$ stands for an at most countable collection of piecewise constant functions on $\mathscr{W}$ with values in J based on the partition m. Setting the notation $\mathbf{\Gamma} = \cup_{m\in\mathcal{M}}\mathbf{\Gamma}_m$, we assume the family of (possibly random) candidates $\widehat{\mathbf{\Gamma}} = \{\widehat{\boldsymbol{\gamma}}_\lambda(\boldsymbol{X}),\ \lambda \in \Lambda\}$ for $\boldsymbol{\gamma}^\star$ (may not exist) with values in $\mathbf{\Gamma}$. This is to say, for each $\lambda \in \Lambda$, there is a (possibly random) partition $\widehat{m}(\lambda) \in \mathcal{M}$

such that $\widehat{\boldsymbol{\gamma}}_\lambda \in \boldsymbol{\Gamma}_{\widehat{m}(\lambda)}$. For any $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}$, we define $\mathcal{M}(\boldsymbol{\gamma}) = \{m \in \mathcal{M},\ \boldsymbol{\gamma} \in \boldsymbol{\Gamma}_m\}$, therefore naturally we have $\widehat{m}(\lambda) \in \mathcal{M}(\widehat{\boldsymbol{\gamma}}_\lambda)$. Let $\Delta(\cdot)$ be a map from $\mathcal{M}$ to $\mathbb{R}_+ = [0, +\infty)$. For each $m \in \mathcal{M}$, we associate it with a nonnegative weight $\Delta(m)$ and assume the following holds true.

Assumption 4.2.1. There exists a positive number Σ such that

$$\Sigma = \sum_{m \in \mathcal{M}} e^{-\Delta(m)} < +\infty. \tag{4.2.1}$$

We remark that when $\Sigma = 1$, the weights $\Delta(m)$ define a prior distribution on the collection of partitions $\mathcal{M}$, which gives a Bayesian flavour to our selection procedure.

Given two partitions $m_1, m_2 \in \mathcal{M}$, we define a refined partition $m_1 \vee m_2$ on $\mathscr{W}$ generated by m_1, m_2 as

$$m_1 \vee m_2 = \{K_1 \cap K_2 \mid K_1 \in m_1,\ K_2 \in m_2,\ K_1 \cap K_2 \neq \varnothing\}.$$

For any partition m on $\mathscr{W}$, we denote the number of its segments by $|m|$. To define our selection procedure, we also make the following assumption on the family $\mathcal{M}$.

Assumption 4.2.2. There exists some constant $\alpha \geq 1$ such that $|m_1 \vee m_2| \leq \alpha(|m_1| + |m_2|)$, for all $m_1, m_2 \in \mathcal{M}$.

We give some examples of the family $\mathcal{M}$ here such that Assumption 4.2.2 holds true. When $\mathscr{W}$ is either $\mathbb{R}$ or some subinterval of $\mathbb{R}$, for any finite or countable family $\mathcal{M}$ of partitions on $\mathscr{W}$, it is easy to observe that Assumption 4.2.2 is satisfied with $\alpha = 1$. Another example can be the nested partitions, i.e. the family $\mathcal{M}$ is ordered for the inclusion. In this situation, $m_1 \vee m_2$ either equals to m_1 or m_2 so that Assumption 4.2.2 also holds true with $\alpha = 1$. Besides, when $\mathscr{W} = [0,1)^d$ with $d \geq 2$, a specific example satisfying Assumption 4.2.2 with $\alpha = 2$ has been introduced in Example 3 of Baraud and Birgé (2009).

Our selection procedure is based on a pair-by-pair comparison of the candidates, where the selection mechanism is inspired by a sequence of work of the ρ-estimation (see Baraud et al. (2017) and Baraud and Birgé (2018)). However, we generalize the comparison device into the situation where the elements in $\widehat{\boldsymbol{\Gamma}}$ can be random. Let us first recall the monotone increasing function ψ from $[0, +\infty]$ into $[-1, 1]$ defined in (2.3.1) as

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

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

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

with the conventions $0/0 = 1$ and $a/0 = +\infty$ for all $a > 0$. Let D_n be a map from $\mathcal{M}$ to $\mathbb{R}_+$ defined as, for any $m \in \mathcal{M}$,

$$D_n(m) = |m| \left[9.11 + \log_+ \left(\frac{n}{|m|} \right) \right],$$

where $\log_+(x) = \max\{\log(x), 0\}$. We define the penalty function from $\mathbf{\Gamma}$ to $\mathbb{R}_+$ such that for all $\boldsymbol{\gamma} \in \mathbf{\Gamma}$,

$$\mathbf{pen}(\boldsymbol{\gamma}) \geq C_0 \left(2\alpha + \frac{1}{2} \right) \inf_{m \in \mathcal{M}(\boldsymbol{\gamma})} \left[D_n(m) + \Delta(m) \right], \tag{4.2.2}$$

where $C_0 > 0$ is a universial constant. For each $\lambda \in \Lambda$, we set

$$\boldsymbol{\upsilon}(\boldsymbol{X}, \widehat{\boldsymbol{\gamma}}_\lambda) = \sup_{\lambda' \in \Lambda} \left[\mathbf{T}(\boldsymbol{X}, \widehat{\boldsymbol{\gamma}}_\lambda, \widehat{\boldsymbol{\gamma}}_{\lambda'}) - \mathbf{pen}(\widehat{\boldsymbol{\gamma}}_{\lambda'}) \right] + \mathbf{pen}(\widehat{\boldsymbol{\gamma}}_\lambda).$$

We select $\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}$ as any measurable element of the random (and non-void) set

$$\mathscr{E}(\boldsymbol{X}) = \left\{ \widehat{\boldsymbol{\gamma}}_\lambda \in \widehat{\mathbf{\Gamma}} \text{ such that } \boldsymbol{\upsilon}(\boldsymbol{X}, \widehat{\boldsymbol{\gamma}}_\lambda) \leq \inf_{\lambda' \in \Lambda} \boldsymbol{\upsilon}(\boldsymbol{X}, \widehat{\boldsymbol{\gamma}}_{\lambda'}) + 1 \right\}. \tag{4.2.3}$$

The final selected estimator $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}}$ of $\mathbf{Q}^\star$ is given by $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}} = \left(R_{\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}}, \ldots, R_{\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}} \right)$.

We comment that the number 1 in (4.2.3) does not play any role, therefore can be substituted by any small number $\delta > 0$. We choose $\delta = 1$ here just for enhancing the legibility of our results. Moreover, to improve the performance of the selected estimator $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}}$, the choice of a $\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}$ such that $\boldsymbol{\upsilon}(\boldsymbol{X}, \widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}) = \inf_{\lambda \in \Lambda} \boldsymbol{\upsilon}(\boldsymbol{X}, \widehat{\boldsymbol{\gamma}}_\lambda)$ should be preferred whenever available, which is the case when $\widehat{\mathbf{\Gamma}}$ is a finite set.

4.2.2 The performance of the selected estimator

In this section, we establish non-asymptotic exponential inequalities of deviations between the selected estimator $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}}$ and $\mathbf{Q}^\star$.

Theorem 4.2.1. *Under Assumption 4.2.1 and 4.2.2, whatever the conditional distributions $\mathbf{Q}^\star = (Q_1^\star, \ldots, Q_n^\star)$ of Y_i given W_i and the distributions of W_i, there exists a universal constant $C_0 > 0$ such that the selected estimator $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}}$ given by the procedure in Section 4.2.1 among a family of (possibly random) candidates $\widehat{\mathbf{\Gamma}} = \{\widehat{\boldsymbol{\gamma}}_\lambda(\boldsymbol{X}),\ \lambda \in \Lambda\}$ based on the observations $\boldsymbol{X} = (X_1, ..., X_n)$ satisfies for any $\xi > 0$, on a set of probability larger than $1 - \Sigma^2 e^{-\xi}$*

$$\mathbf{h}^2(\mathbf{Q}^\star, \mathbf{R}_{\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}}) \leq \inf_{\lambda \in \Lambda} \left[c_1 \mathbf{h}^2(\mathbf{Q}^\star, \mathbf{R}_{\widehat{\boldsymbol{\gamma}}_\lambda}) + c_2 \mathbf{pen}(\widehat{\boldsymbol{\gamma}}_\lambda) \right] + c_3 \left(1.471 + \xi \right), \tag{4.2.4}$$

where $c_1 = 91.4$, $c_2 = 42.7$ and $c_3 = 12666.9$.

The proof of Theorem 4.2.1 is postponed to Section 4.6. We hereby give a short discussion for the numerical constant C_0 in the penalty function (4.2.2). In the proof

of Theorem 4.2.1, we show that there does exist a numerical constant $C_0 > 0$ such that for all the penalties satisfying (4.2.2), the procedure defined in Section 4.2.1 results in a selected estimator fulfilling the performance stated in Theorem 4.2.1. Unfortunately, this theoretical constant C_0 turns out to be quite large and we do not have enough information about the smallest value of C_0 which validates the non-asymptotic exponential inequalities in (4.2.4). In practice, when we implement our estimator selection procedure we regard this C_0 as a tuning parameter instead of using the theoretical value. For this point, we will make it more clear in the simulation study, where it also turns out the value of C_0 in theory seems to be too pessimistic.

To comment on the performance of the selected estimator further, we integrate (4.2.4) with respect to ξ and obtain the following risk bound.

Corollary 4.2.1. *Under Assumption 4.2.1 and 4.2.2, whatever the conditional distributions* $\mathbf{Q}^\star = (Q_1^\star, \ldots, Q_n^\star)$ *of* Y_i *given* W_i *and the distributions of* W_i*, there exists a universal constant* $C_0 > 0$ *such that the selected estimator* $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}}$ *given by the procedure in Section 4.2.1 among* $\widehat{\boldsymbol{\Gamma}} = \{\widehat{\boldsymbol{\gamma}}_\lambda(\boldsymbol{X}),\ \lambda \in \Lambda\}$ *satisfies*

$$\begin{aligned}\mathbb{E}\left[\mathbf{h}^2(\mathbf{Q}^\star, \mathbf{R}_{\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}})\right] &\le \mathbb{E}\left[\inf_{\lambda\in\Lambda}\left(c_1\mathbf{h}^2(\mathbf{Q}^\star, \mathbf{R}_{\widehat{\boldsymbol{\gamma}}_\lambda}) + c_2\mathbf{pen}(\widehat{\boldsymbol{\gamma}}_\lambda)\right)\right] + c_3\left(\Sigma^2 + 1.471\right)\\ &\le \inf_{\lambda\in\Lambda}\left\{\mathbb{E}\left[c_1\mathbf{h}^2(\mathbf{Q}^\star, \mathbf{R}_{\widehat{\boldsymbol{\gamma}}_\lambda}) + c_2\mathbf{pen}(\widehat{\boldsymbol{\gamma}}_\lambda)\right]\right\} + c_3\left(\Sigma^2 + 1.471\right).\end{aligned}$$

In particular, if the equality in (4.2.2) holds,

$$\mathbb{E}\left[\mathbf{h}^2(\mathbf{Q}^\star, \mathbf{R}_{\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}})\right] \le C_{\alpha,\Sigma}\inf_{\lambda\in\Lambda}\left\{\mathbb{E}\left[\mathbf{h}^2(\mathbf{Q}^\star, \mathbf{R}_{\widehat{\boldsymbol{\gamma}}_\lambda})\right] + \mathbb{E}\left[\boldsymbol{\Xi}(\widehat{\boldsymbol{\gamma}}_\lambda)\right]\right\}, \tag{4.2.5}$$

where for all $\lambda \in \Lambda$*,*

$$\begin{aligned}\boldsymbol{\Xi}(\widehat{\boldsymbol{\gamma}}_\lambda) &= \inf_{m\in\mathcal{M}(\widehat{\boldsymbol{\gamma}}_\lambda)}\left[|m|\left(9.11 + \log_+\left(\frac{n}{|m|}\right)\right) + \Delta(m)\right]\\ &\le |\widehat{m}(\lambda)|\left[9.11 + \log_+\left(\frac{n}{|\widehat{m}(\lambda)|}\right)\right] + \Delta(\widehat{m}(\lambda))\end{aligned}$$

and

$$C_{\alpha,\Sigma} = \left[c_2C_0\left(2\alpha + \frac{1}{2}\right) + \frac{c_3\left(\Sigma^2 + 1.471\right)}{9.11}\right] \vee c_1.$$

The result in (4.2.5) compares the risk of the selected estimator $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}}$ to those of $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}_\lambda}$ plus an additional nonnegative term $\mathbb{E}\left[\boldsymbol{\Xi}(\widehat{\boldsymbol{\gamma}}_\lambda)\right]$. One nice feature of this approach implied by (4.2.5) lies in the fact that the risk bound does not depend on the cardinality of the set $\widehat{\boldsymbol{\Gamma}}$. This entails that if we enlarge the collection of our candidates by keeping $\mathcal{M}$ unchanged (so that $\Delta(m)$ will not change), the risk bound for the selected estimator only decreases over the larger collection of candidates. On the other hand, our procedure is based on $\mathcal{O}(|\widehat{\boldsymbol{\Gamma}}|^2)$ times of pair-by-pair comparisons. Therefore, the payment for enlarging set $\widehat{\boldsymbol{\Gamma}}$ is the computation time.

The risk bound (4.2.5) in Corollary 4.2.1 also accounts for the stability of our selection procedure under a slight misspecification framework. To illustrate, let us first consider the ideal situation where $\mathbf{Q}^\star = \mathbf{R}_{\boldsymbol{\gamma}^\star} = (R_{\boldsymbol{\gamma}^\star}, \ldots, R_{\boldsymbol{\gamma}^\star})$ with $\boldsymbol{\gamma}^\star$ a piecewise constant function based on the partition $m^\star$ of $\mathscr{W}$. We denote $\overline{\boldsymbol{\Gamma}}_{m^\star}$ the class of all piecewise constant functions with values in $J \subset \mathbb{R}$ based on the partition $m^\star$ and assume for simplicity $\widehat{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}_{m^\star}$, where $\boldsymbol{\Gamma}_{m^\star}$ stands for a dense (for the topology of the pointwise convergence) and countable subset of $\overline{\boldsymbol{\Gamma}}_{m^\star}$. Taking $\Delta(m^\star) = 0$, we deduce from (4.2.5) that the estimator $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}}$ based on the selection among $\boldsymbol{\Gamma}_{m^\star}$ satisfies

$$\mathbb{E}\left[\mathbf{h}^2(\mathbf{R}_{\boldsymbol{\gamma}^\star}, \mathbf{R}_{\widehat{\boldsymbol{\gamma}}})\right] \leq C|m^\star| \left[1 + \log_+\left(\frac{n}{|m^\star|}\right)\right], \tag{4.2.6}$$

where C is a numerical constant. As we have seen in Chapter 2, up to a logarithm term, the right hand side of (4.2.6) is of the expected order of magnitude $|m^\star|$ for the quantity $\mathbf{h}^2(\mathbf{R}_{\boldsymbol{\gamma}^\star}, \mathbf{R}_{\widehat{\boldsymbol{\gamma}}})$. If it is not the ideal case, an approximation error $\mathbf{h}^2(\mathbf{Q}^\star, \boldsymbol{\mathscr{Q}}_{m^\star})$ with $\boldsymbol{\mathscr{Q}}_{m^\star} = \{\mathbf{R}_{\boldsymbol{\gamma}},\ \boldsymbol{\gamma} \in \boldsymbol{\Gamma}_{m^\star}\}$, will be added into the right hand side of (4.2.6) according to (4.2.5). However, as long as this bias term remains small, the performance of our selected estimator will not deteriorate too much as compared to the ideal situation.

4.2.3 Connection to model selection

The work done in this chapter differs from the corresponding result (3.2.7) given by model selection procedure in Chapter 3. In fact, one can regard Corollary 4.2.1 as a more general result of the one in Chapter 3. We illustrate this connection as follows.

We consider the particular application of our selection procedure in the context of model selection. For simplicity, let the equality holds in (4.2.2). We take $\Lambda = \{1, \ldots, |\boldsymbol{\Gamma}|\}$ which is the index set of all the functions belonging to $\boldsymbol{\Gamma} = \cup_{m \in \mathcal{M}} \boldsymbol{\Gamma}_m$ so that in this case, $\widehat{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma} = \{\boldsymbol{\gamma}_\lambda,\ \lambda \in \Lambda\}$ is a collection of deterministic candidates. Moreover, for each $\lambda \in \Lambda$, there exists a deterministic $m(\lambda) \in \mathcal{M}$ such that $\boldsymbol{\gamma}_\lambda \in \boldsymbol{\Gamma}_{m(\lambda)}$. Let us denote $\boldsymbol{\mathscr{Q}}_m = \{\mathbf{R}_{\boldsymbol{\gamma}},\ \boldsymbol{\gamma} \in \boldsymbol{\Gamma}_m\}$, for each $m \in \mathcal{M}$. We can immediately deduce from (4.2.5) that the estimator $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}}$ based on the selection among the family $\{\boldsymbol{\Gamma}_m,\ m \in \mathcal{M}\}$ satisfies

$$\mathbb{E}\left[\mathbf{h}^2(\mathbf{Q}^\star, \mathbf{R}_{\widehat{\boldsymbol{\gamma}}})\right] \leq C_{\alpha,\Sigma} \inf_{m \in \mathcal{M}} \left[\mathbf{h}^2(\mathbf{Q}^\star, \boldsymbol{\mathscr{Q}}_m) + D_n(m) + \Delta(m)\right],$$

which is, up to constants, the result (3.2.7) in Chapter 3 when one takes $\boldsymbol{\Gamma}_m$ as the collection of piecewise constant functions on $\mathscr{W}$. The difference is the model selection procedure, on the one hand, does not require Assumption 4.2.2 to be satisfied and can be applied to other types of models to approximate the potential $\boldsymbol{\gamma}^\star$ besides piecewise constant ones. On the other hand, when the number of models becomes large, model selection strategy is more of theoretical interest due to its expensive numerical cost. The estimator selection strategy, however, allows to deal with random partitions which can be

obtained for example from dynamic programming algorithm (see Rigaill (2015)) or CART algorithm (see Breiman et al. (1984)). Efficiently reducing the cardinality of $\widehat{\mathbf{\Gamma}}$, these algorithms together with our estimator selection procedure take the model selection strategy into practice. Moreover, the idea that selecting among random candidates set makes the selection between estimators given by different model selection strategies possible.

4.3 Application to changepoint detection in exponential families

In this section, we consider the application of our estimator selection procedure to changepoint detection problem in exponential families. In such a context, people usually assume the exponential family $\overline{\mathscr{Q}} = \{R_\gamma,\ \gamma \in J\}$ has been parametrized in its natural form which entails u is taken as the identity function in (2.3.12). We observe a sequence $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ with values in $\mathscr{Y}^n$ and assume that there exists a vector $\gamma^\star = (\gamma_1^\star, \ldots, \gamma_n^\star) \in J^n$ with $N-1$ changepoints, $N \geq 1$ such that within each segment, the values of $\gamma^\star$ remain a constant and for each $i \in \{1, \ldots, n\}$, the distribution of Y_i is given by $R_{\gamma_i^\star}$. This corresponds to the situation in our setting when $W_i = (i-1)/n$ are deterministic, for all $i \in \{1, \ldots, n\}$ so that $\mathscr{W} = [0, 1)$ and the function $\boldsymbol{\gamma}^\star : [0, 1) \to J \subset \mathbb{R}$ is a right-continuous step function with $N \geq 1$ segments. For a consistency with the former paragraphs, we take $W_i = (i-1)/n$ throughout this section and use the function notation $\boldsymbol{\gamma}^\star$ rather than the vector $\gamma^\star \in J^n$ in the sequel.

For each $1 \leq k \leq n$, let $\mathcal{M}_k$ stand for the collection of all possible partitions of the sequence $1, \ldots, n$ into k segments and denote $\mathcal{M} = \cup_{1 \leq k \leq n} \mathcal{M}_k$. In changepoint detection problem, for each $m \in \mathcal{M}$, we assign its weight as

$$\Delta(m) = \log \binom{n-1}{|m|-1} + |m|. \tag{4.3.1}$$

With (4.3.1), a basic computation leads to $\Sigma = \sum_{m \in \mathcal{M}} \exp\left[-\Delta(m)\right] \leq 1/(e-1)$ which entails Assumption 4.2.1 is satisfied. Moreover, since $\mathscr{W} = [0, 1) \subset \mathbb{R}$, for any $m_1, m_2 \in \mathcal{M}$, $|m_1 \vee m_2| \leq |m_1| + |m_2| - 1$, Assumption 4.2.2 also holds true with $\alpha = 1$.

Supposing that we have a finite collection of (possibly random) piecewise constant candidates $\widehat{\mathbf{\Gamma}} = \{\widehat{\boldsymbol{\gamma}}_\lambda(\boldsymbol{X}),\ \lambda \in \Lambda\}$, we associate each $\widehat{\boldsymbol{\gamma}}_\lambda(\boldsymbol{X})$ with the penalty

$$\mathbf{pen}(\widehat{\boldsymbol{\gamma}}_\lambda) = \kappa \left\{ |\widehat{m}(\lambda)| \left[10.11 + \log\left(\frac{n}{|\widehat{m}(\lambda)|} \right) \right] + \log \binom{n-1}{|\widehat{m}(\lambda)|-1} \right\},$$

where $\kappa = 2.5 C_0$ is the parameter to be tuned later. Once the value of κ is given, our estimator selection procedure can be implemented by running Algorithm 2.

Algorithm 2 Estimator selection

Input:

$\boldsymbol{X} = (X_1, \cdots, X_n)$: the observations.

Output: $\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}$

1: Collect $\widehat{\boldsymbol{\Gamma}} = \{\widehat{\boldsymbol{\gamma}}_\lambda,\ \lambda \in \Lambda\}$ based on $\boldsymbol{X}$.
2: **for** $\lambda \in \Lambda$ **do**
3: $\quad \boldsymbol{\upsilon}(\boldsymbol{X}, \widehat{\boldsymbol{\gamma}}_\lambda) \leftarrow \sup_{\lambda' \in \Lambda} \left[\mathbf{T}(\boldsymbol{X}, \widehat{\boldsymbol{\gamma}}_\lambda, \widehat{\boldsymbol{\gamma}}_{\lambda'}) - \mathbf{pen}(\widehat{\boldsymbol{\gamma}}_{\lambda'})\right] + \mathbf{pen}(\widehat{\boldsymbol{\gamma}}_\lambda)$.
4: **end for**
5: $\widehat{\lambda} \leftarrow \operatorname{argmin}_{\lambda \in \Lambda} \boldsymbol{\upsilon}(\boldsymbol{X}, \widehat{\boldsymbol{\gamma}}_\lambda)$.
6: Return $\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}$.

4.3.1 Calibrating the value of κ

We take $\kappa = 0.08$ uniformly over all the exponential families. The reason for this choice of κ is explained in this section.

The idea to calibrate the value of κ is rather simple. Roughly speaking, we first simulate data of size n and prepare a collection of candidates $\widehat{\boldsymbol{\Gamma}}$ which can be done by running the algorithm in R package `Segmentor3IsBack` (implementing the procedure proposed by Cleynen and Lebarbier (2014, 2017)). Then we take different values of κ to design our penalty and obtain a sequence of the selected $\widehat{\boldsymbol{\gamma}}_{\kappa,\widehat{\lambda}}$ among $\widehat{\boldsymbol{\Gamma}}$ associated to various κ. For each value of κ, we repeat the experiment in each simulation setting 100 times and finally evaluate the risk $\mathbb{E}\left[\mathbf{h}^2\left(\mathbf{Q}^\star, \mathbf{R}_{\widehat{\boldsymbol{\gamma}}_{\kappa,\widehat{\lambda}}}\right)\right]$ of the selected estimator $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}_{\kappa,\widehat{\lambda}}}$ by its empirical mean, namely we compute

$$\widehat{R}_n\left(\widehat{\boldsymbol{\gamma}}_{\kappa,\widehat{\lambda}}\right) = \frac{1}{100}\sum_{l=1}^{100}\left[\sum_{i=1}^{n} h^2\left(Q_i^\star, R_{\widehat{\boldsymbol{\gamma}}^l_{\kappa,\widehat{\lambda}}\left(\frac{i-1}{n}\right)}\right)\right],$$

where $\widehat{\boldsymbol{\gamma}}^l_{\kappa,\widehat{\lambda}}$ is the l-th realisation of the selected estimator associated to a fixed κ.

Simulating data

The experiments have been done for three models: Gaussian, Poisson and exponential changepoint detection.

Let $\boldsymbol{\gamma}^\star$ be piecewise constant on $[0, 1)$ with N segments and $\mathbf{Q}^\star = \mathbf{R}_{\boldsymbol{\gamma}^\star}$. For each model, we design the experiments under three settings where for all the settings $n = 500$, but $N = 5$, $N = 10$ and $N = 20$ respectively. For all the three settings, the changepoints are uniformly located, i.e. every 100 data-points for the first setting, every 50 data-points in the second setting and every 25 data-points in the third setting.

— Under all the settings of Gaussian model, for $1 \leq i \leq n$, if Y_i locates at the j-th segment with $1 \leq j \leq N$, Y_i follows a Gaussian distribution with mean $(j+1)/2$, variance $\sigma^2 = 1$.

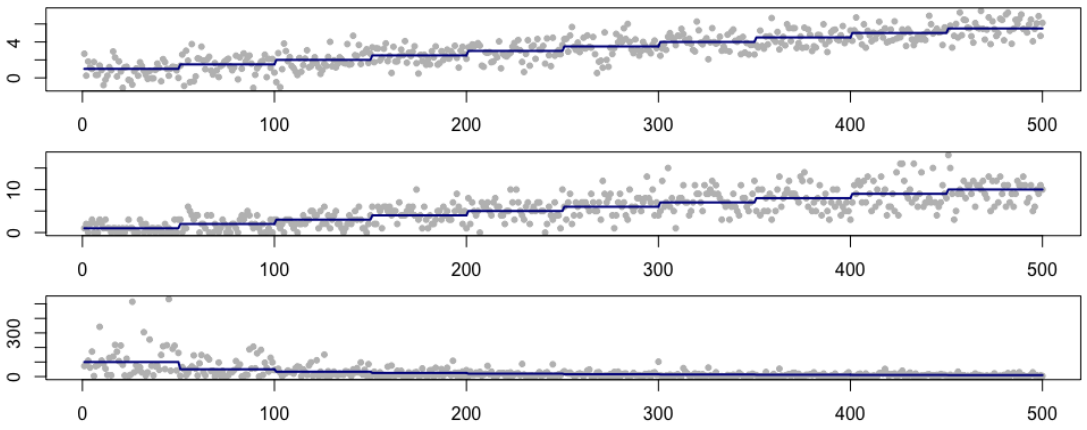


Figure 4.1: The 1st graph (top) corresponds to one profile of the simulated data (dots) and $\boldsymbol{\gamma}^\star$ (solid line) for Gaussian model; The 2nd graph (middle) corresponds to one profile of the simulated data (dots) and $\exp(\boldsymbol{\gamma}^\star)$ (solid line) for Poisson model; The 3rd graph (bottom) corresponds to one profile of the simulated data (dots) and $1/\boldsymbol{\gamma}^\star$ (solid line) for exponential model.

— Under all the settings of Poisson model, for $1 \leq i \leq n$, if Y_i locates at the j-th segment with $1 \leq j \leq N$, Y_i follows a Poisson distribution with mean j which means $\boldsymbol{\gamma}^\star$ takes value $\log(j)$ on the j-th segment.

— Under all the settings of exponential model, for $1 \leq i \leq n$, if Y_i locates at the j-th segment with $1 \leq j \leq N$, Y_i follows an exponential distribution with natural parameter $0.01j$.

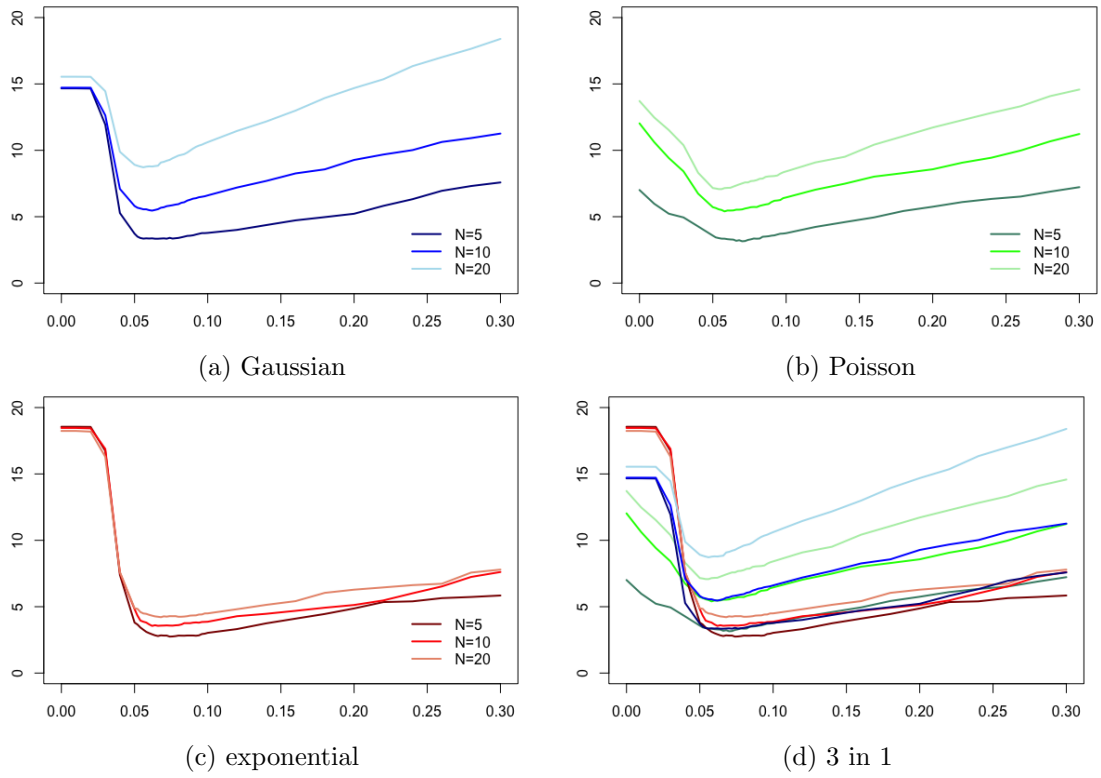
Figure 4.1 exhibits one example of the simulated data (when $N = 10$) and the true value of the regression function $\boldsymbol{\gamma}^\star$ (or a suitable transformation of $\boldsymbol{\gamma}^\star$) on each segment.

Collecting candidates in $\widehat{\mathbf{\Gamma}}$

In the work of Cleynen and Lebarbier (2014, 2017), they solved this problem by a model selection procedure via some suitable penalty function based on the partitions given by the pruned dynamic programming algorithm (PDPA for short) proposed by Rigaill (2015). Given $N_{\max}$ the maximum number of segments for consideration, for each integer λ with $1 \leq \lambda \leq N_{\max}$, PDPA searches the optimal partition with exact λ segments. We set $N_{\max} = 30$ hence 30 partitions of the sequence $1, \ldots, n$ are returned by PDPA. Provided a partition, the value of $\boldsymbol{\gamma}^\star$ on each segment is given by MLE as in Cleynen and Lebarbier (2014, 2017). By doing so, we collect 30 candidates which we denote as $\widehat{\mathbf{\Gamma}}_c = \{\widehat{\boldsymbol{\gamma}}_\lambda, 1 \leq \lambda \leq N_{\max}\}$.

Results

Under each setting of all the three models, one experiment means we simulate $n = 500$ observations with N segments based on the corresponding $\boldsymbol{\gamma}^\star$ introduced in Section 4.3.1. We then select the estimator among the candidate ones $\widehat{\boldsymbol{\Gamma}}_c$ by Algorithm 2 via the penalty functions associated to different values of κ. Finally we observe the quantity $\widehat{R}_n(\widehat{\boldsymbol{\gamma}}_{\kappa,\widehat{\lambda}})$ and regard it as the criterion to calibrate a suitable value of κ. The results for all nine settings are shown in Figure 4.2, where the horizontal axis represents the values of κ and the vertical indicates the quantity $\widehat{R}_n(\widehat{\boldsymbol{\gamma}}_{\kappa,\widehat{\lambda}})$.



(a) Gaussian (b) Poisson

(c) exponential (d) 3 in 1

Figure 4.2: $\widehat{R}_n(\widehat{\boldsymbol{\gamma}}_{\kappa,\widehat{\lambda}})$ with respect to κ under nine settings.

In Figure 4.2, the quantities $\widehat{R}_n(\widehat{\boldsymbol{\gamma}}_{\kappa,\widehat{\lambda}})$ under all nine settings have a tendency to first decrease and then increase with respect to the increasing of κ, which is consistent to the theoretical results. When κ is too small, the penalty function is relatively small for the complexed models therefore the overfitting issue may happen. However, when κ is too large, the penalty function is excessively large for the complexed models which will cause an overpenalization. Moreover, the minimizers of κ for the quantities $\widehat{R}_n(\widehat{\boldsymbol{\gamma}}_{\kappa,\widehat{\lambda}})$ in all nine settings are close to each other and all concentrate within a short interval $[0.05, 0.1]$. From the slope in all nine settings, we observe that overfitting issue will cause

a more seriously negative influence to the accuracy than overpenalization. Considering the optimal performance of all the settings and also being safe with respect to overfitting, we choose κ as the largest minimizer of $\widehat{R}_n(\widehat{\gamma}_{\kappa,\widehat{\lambda}})$ among nine settings which approximately equals to 0.08 and implement our procedure with $\kappa = 0.08$ in later studies.

4.4 Simulation study and discussion

Throughout this section, we carry out a comparative simulation study with the state-of-art competitors available in R packages for changepoint detection problem in exponential families. Unless otherwise specified, the competitors are implemented under the default settings in their packages. For Gaussian model, some of our competitors use the estimated value of the standard deviation σ. To make the comparison as fair as possible, we also implement the median absolute deviation estimator for σ while running our procedure, which is the one adopted in Killick et al. (2012) and Fearnhead and Rigaill (2019).

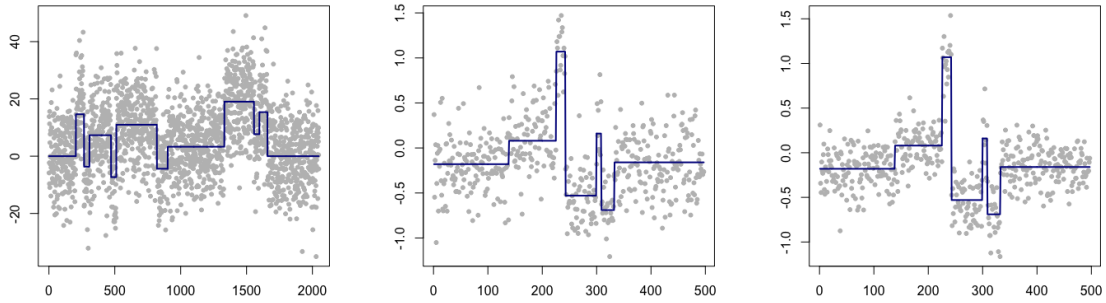
To evaluate the performance of an estimator, besides the empirical risk $\widehat{R}_n(\cdot)$ obtained from replications, we also record $\widehat{N} - N$ which computes the difference between the estimated number of segments and the truth for each replication.

4.4.1 Accuracy

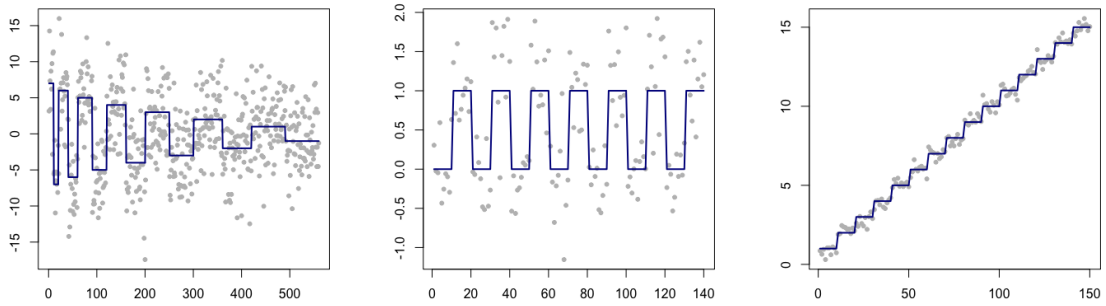
In this section, we study the changepoint detection problem for Gaussian model where numerous literature can be found tackling this issue. We construct our candidates set $\widehat{\mathbf{\Gamma}}$ as a collection of some cutting-edge estimators with implemented R packages and these ones are also regarded as the competitors of our estimator ES. More precisely, the competing packages we consider are: `PSCBS`, which implements the CBS procedure proposed in Olshen et al. (2004); `cumSeg`, which performs the method given by Muggeo and Adelfio (2010); `changepoint`, which implements the PELT approach provided by Killick et al. (2012); `StepR`, which implements the SMUCE given by Frick et al. (2013); `Segmentor3IsBack`, which implements CL proposed by Cleynen and Lebarbier (2014, 2017); `wbs`, which implements the wild binary segmentation methodology proposed in Fryźlewicz (2014); `FDRSeg`, which implements the approach given in Li et al. (2016); `robseg`, which implements the procedure proposed by Fearnhead and Rigaill (2019). We would like to study the performance of our estimator ES based on the selection among these state-of-art ones.

We follow the test signals considered by Fryźlewicz (2014) and then by Fearnhead and Rigaill (2019) which involves 5 different formats of signals with length from $n = 140$ to 2048: (1) `blocks`, (2) `fms`, (3) `mix`, (4) `teeth10` and (5) `stairs10`. The specific settings of these signals including the sample sizes and noise standard deviations are given in Appendix B of Fryźlewicz (2014). Following the experiments done in Fearnhead and Rigaill (2019), we also consider an additional signal setting by changing the standard

deviation of (2) `fms` from 0.3 into 0.2, which is also one of the settings studied in Frick et al. (2013). An example of one profile of the simulated data and the underlying signals $\boldsymbol{\gamma}^\star$ are plotted in Figure 4.3. For each signal, the experiment has been replicated 1000 times. The results are shown in Table 4.1. The performance of each estimator is stated as follows.



(a) blocks (b) fms ($\sigma = 0.3$) (c) fms ($\sigma = 0.2$)



(d) mix (e) teeth10 (f) stairs10

Figure 4.3: The six signals (solid line) and simulated data (dots).

CBS and cumSeg. The CBS and cumSeg in general behave poorly compared with other procedures. The CBS only has satisfactory performance of detecting changes for `blocks` and `fms` ($\sigma = 0.2$) but it turns out CBS always results in a relatively large $\widehat{R}_n(\cdot)$. Except acceptable performance for `fms` ($\sigma = 0.2$) and `stairs10`, cumSeg always tends to underestimate the number of changes and also yields an estimator with quite large empirical risk.

PELT. The PELT has excellent performance for both of the `fms` signals and `stairs10`. For `blocks` signal, it is above the average but does not belong to the first class among all. As for `mix` and `teeth10`, it performs rather average.

SMUCE. The SMUCE has very excellent performance for `fms` ($\sigma = 0.2$). However, it behaves poorly for all the other signals.

CL. The CL has nice performance for `teeth10`. For `blocks` and `mix`, its performance is satisfactory though not belonging to the first class. For both of the `fms` signals, it

		$\widehat{N}-N$						
Method	Signal	≤ -2	-1	0	1	≥ 2	$\widehat{R}_n(\cdot)$	Contribution
ES	blocks	0.005	0.278	**0.656**	0.055	0.006	5.61 ± 0.12	-
CBS	blocks	0.006	0.090	0.575	0.184	0.145	7.57 ± 0.14	0.000
cumSeg	blocks	0.653	0.335	0.011	0.001	0.000	15.71 ± 0.40	0.000
PELT	blocks	0.014	0.389	0.574	0.020	0.003	5.69 ± 0.11	0.035
SMUCE	blocks	0.940	0.060	0.000	0.000	0.000	16.02 ± 0.37	0.010
CL	blocks	0.010	0.356	0.595	0.035	0.004	5.67 ± 0.12	0.533
WBS sSIC	blocks	0.021	0.412	0.532	0.032	0.003	6.11 ± 0.13	0.013
FDR($\alpha = 0.05$)	blocks	0.008	0.447	0.478	0.059	0.008	6.15 ± 0.13	0.332
robseg(Huber)	blocks	0.004	0.234	**0.674**	0.072	0.016	5.84 ± 0.12	0.063
robseg(biweight)	blocks	0.020	0.404	0.558	0.017	0.001	5.88 ± 0.12	0.014
ES	fms(0.3)	0.008	0.002	**0.915**	0.069	0.006	2.16 ± 0.07	-
CBS	fms(0.3)	0.007	0.012	0.796	0.139	0.046	5.10 ± 0.09	0.000
cumSeg	fms(0.3)	0.706	0.041	0.224	0.028	0.001	7.07 ± 0.44	0.000
PELT	fms(0.3)	0.007	0.003	**0.922**	0.061	0.007	2.15 ± 0.08	0.054
SMUCE	fms(0.3)	0.074	0.537	0.388	0.001	0.000	5.15 ± 0.18	0.293
CL	fms(0.3)	0.002	0.001	0.837	0.119	0.041	2.28 ± 0.08	0.199
WBS sSIC	fms(0.3)	0.007	0.003	**0.933**	0.048	0.009	2.26 ± 0.08	0.008
FDR($\alpha = 0.05$)	fms(0.3)	0.001	0.027	**0.879**	0.076	0.017	2.28 ± 0.09	0.409
robseg(Huber)	fms(0.3)	0.001	0.001	0.825	0.130	0.043	2.37 ± 0.08	0.007
robseg(biweight)	fms(0.3)	0.013	0.005	**0.928**	0.049	0.005	2.23 ± 0.08	0.030
ES	fms(0.2)	0.000	0.000	**0.923**	0.071	0.006	1.61 ± 0.06	-
CBS	fms(0.2)	0.000	0.000	0.871	0.086	0.043	5.79 ± 0.07	0.000
cumSeg	fms(0.2)	0.094	0.009	0.812	0.083	0.002	5.19 ± 0.22	0.002
PELT	fms(0.2)	0.000	0.000	**0.929**	0.060	0.011	1.59 ± 0.06	0.022
SMUCE	fms(0.2)	0.000	0.001	**0.994**	0.005	0.000	1.49 ± 0.06	0.734
CL	fms(0.2)	0.000	0.000	0.840	0.128	0.032	1.74 ± 0.07	0.102
WBS sSIC	fms(0.2)	0.000	0.000	**0.945**	0.050	0.005	1.65 ± 0.06	0.003
FDR($\alpha = 0.05$)	fms(0.2)	0.000	0.000	0.871	0.103	0.026	1.66 ± 0.06	0.115
robseg(Huber)	fms(0.2)	0.000	0.000	0.830	0.135	0.035	1.83 ± 0.07	0.008
robseg(biweight)	fms(0.2)	0.000	0.000	**0.937**	0.058	0.005	1.63 ± 0.06	0.014
ES	mix	0.264	0.243	**0.434**	0.056	0.003	5.91 ± 0.12	-
CBS	mix	0.313	0.201	0.324	0.109	0.053	11.18 ± 0.17	0.000
cumSeg	mix	0.999	0.001	0.000	0.000	0.000	32.61 ± 0.92	0.000
PELT	mix	0.375	0.270	0.321	0.032	0.002	6.11 ± 0.12	0.070
SMUCE	mix	0.922	0.076	0.002	0.000	0.000	12.59 ± 0.42	0.042
CL	mix	0.305	0.244	0.390	0.053	0.008	6.04 ± 0.12	0.585
WBS sSIC	mix	0.342	0.269	0.351	0.032	0.006	5.99 ± 0.12	0.029
FDR($\alpha = 0.05$)	mix	0.411	0.358	0.181	0.038	0.012	6.71 ± 0.13	0.190
robseg(Huber)	mix	0.209	0.240	**0.444**	0.088	0.019	6.10 ± 0.12	0.051
robseg(biweight)	mix	0.403	0.264	0.305	0.026	0.002	6.30 ± 0.12	0.033
ES	teeth10	0.215	0.025	**0.721**	0.037	0.002	5.69 ± 0.24	-
CBS	teeth10	0.999	0.000	0.001	0.000	0.000	24.69 ± 0.07	0.000
cumSeg	teeth10	1.000	0.000	0.000	0.000	0.000	24.85 ± 0.01	0.005
PELT	teeth10	0.274	0.029	0.657	0.037	0.003	6.03 ± 0.24	0.090
SMUCE	teeth10	0.984	0.013	0.003	0.000	0.000	20.11 ± 0.22	0.003
CL	teeth10	0.029	0.013	**0.679**	0.204	0.075	4.71 ± 0.13	0.321
WBS sSIC	teeth10	0.067	0.021	**0.752**	0.120	0.040	5.30 ± 0.26	0.010
FDR($\alpha = 0.05$)	teeth10	0.309	0.135	0.508	0.040	0.008	7.68 ± 0.32	0.356
robseg(Huber)	teeth10	0.105	0.026	**0.748**	0.102	0.019	4.94 ± 0.15	0.016
robseg(biweight)	teeth10	0.318	0.028	0.635	0.019	0.000	6.31 ± 0.25	0.199
ES	stairs10	0.00	0.004	**0.949**	0.044	0.003	3.33 ± 0.09	-
CBS	stairs10	0.012	0.172	0.789	0.027	0.000	13.81 ± 0.16	0.000
cumSeg	stairs10	0.024	0.090	0.819	0.067	0.000	8.61 ± 0.24	0.000
PELT	stairs10	0.000	0.004	**0.955**	0.039	0.002	3.32 ± 0.09	0.017
SMUCE	stairs10	0.801	0.137	0.062	0.000	0.000	22.26 ± 0.58	0.050
CL	stairs10	0.000	0.001	0.768	0.184	0.047	3.50 ± 0.09	0.178
WBS sSIC	stairs10	0.000	0.001	0.608	0.301	0.090	3.91 ± 0.10	0.004
FDR($\alpha = 0.05$)	stairs10	0.002	0.028	**0.896**	0.053	0.021	3.57 ± 0.12	0.703
robseg(Huber)	stairs10	0.000	0.000	0.867	0.110	0.023	3.45 ± 0.09	0.006
robseg(biweight)	stairs10	0.000	0.005	**0.964**	0.031	0.000	3.36 ± 0.09	0.042

Table 4.1: Frequencies of $\widehat{N} - N$ and $\widehat{R}_n(\cdot)$ of ES and its competitors for Gaussian model over 1000 simulated sample paths. Contribution denotes the frequency of each competitor being selected as ES. Bold: highest empirical frequency of $\widehat{N} - N = 0$ and those with frequencies within 10% off the highest. The uncertainty is obtained by computing $2\widehat{\sigma}/\sqrt{n_r}$, where $\widehat{\sigma}^2$ is the empirical variance and n_r is the number of replications.

shows rather average performance. The CL does not behave well for the `stairs10` signal where it tends to overestimate the number of changes compared to other methods. Let us remark here that the performance of CL in our simulation study is better than the corresponding context in Fryźlewicz (2014). This is because when implementing the package `Segmentor3IsBack`, users need to set the maximum number of segments $N_{\max}$. We set $N_{\max} = 20$ for all the six signals considering the maximal number of changepoints (i.e. $N-1$) among six signals is 14 and they set $N_{\max} = 15$ which resulted in a systematical underestimation of the number of changepoints for CL in their study.

WBS sSIC. We implement the package `wbs` combining the WBS method with the sSIC stopping criterion which, as it has been shown in Fryźlewicz (2014), is the overall winner compared to combining the WBS method with other thresholding stopping rules. The WBS sSIC has excellent performance for both of the `fms` signals and `teeth10`. However, it performs rather average for `blocks` and `mix`. As for `stairs10`, the performance of WBS sSIC is a little poor as a consequence of overestimating the number of changepoints. Such a result has also been confirmed by the study of WBS sSIC in Fryźlewicz (2014).

FDR. The FDR with $\alpha = 0.05$ performs well for `fms` ($\sigma = 0.3$) and `satirs10` signals. For `fms` ($\sigma = 0.2$), it has an average performance. But it behaves below the average under other test signals.

robseg. We consider Huber loss and biweight loss when implementing the package `robseg` which are the recommended ones (especially the biweight loss) according to Fearnhead and Rigaill (2019). The robseg (Huber) performs excellently for `blocks`, `mix` and `teeth10`. It behaves rather average for both of the `fms` signals and `stairs10`. The robseg (biweight) performs excellently for both of the `fms` signals and `stairs10`. As for `blocks`, `mix` and `teeth10` signals, it performs rather average.

ES. As we can observe from the column named "Contribution" in Table 4.1, under different test signals, our estimator selection procedure tends to allocate different preference to the candidates in $\widehat{\mathbf{\Gamma}}$ based on their practical performance. For example, when SMUCE shows obvious outperformance for the signal `fms` ($\sigma = 0.2$), we select it with a frequency 0.734 as our ES estimator. However, we automatically reduce the frequency to select SMUCE as ES when it performs poorly under other signals but prefer some more competitive ones. As a final result, the ES estimator shows a very competitive performance under all the test signals. The interesting point is that this cannot be achieved by any single candidate in $\widehat{\mathbf{\Gamma}}$ since as we have seen above, each of them only outperforms others for some of the test signals but not all.

4.4.2 Stability when outliers present

As we have mentioned in the theoretical analysis part, our estimator selection procedure possesses the stability when there is a slight departure from the presumption $\mathbf{Q}^\star = \mathbf{R}_{\boldsymbol{\gamma}^\star}$

with $\boldsymbol{\gamma}^\star$ being piecewise constant on $\mathscr{W}$. One of the application scenario for this property is when there is a small amount of outliers in the observations which has attracted more attention recently in the changepoint detection. In this section, we test the practical performance of ES as well as its competitors when outliers present. We take the signal `fms` ($\sigma = 0.2$) as an example since most of the existing methods behave rather well under this signal. Based on this signal, we add outliers by randomly choosing five points among the sequence of length $n = 497$ and modifying the values of them into 3. The results of all the estimators are shown in Table 4.2.

			$\widehat{N} - N$						
Method	Signal	Outlier	≤ -2	-1	0	1	≥ 2	$\widehat{R}_n(\cdot)$	Contribution
ES	fms(0.2)	Yes	0.000	0.000	**0.956**	0.043	0.001	1.64 ± 0.06	-
CBS	fms(0.2)	Yes	0.660	0.282	0.038	0.016	0.004	34.55 ± 0.79	0.000
cumSeg	fms(0.2)	Yes	0.801	0.056	0.083	0.021	0.039	16.96 ± 0.51	0.000
PELT	fms(0.2)	Yes	0.000	0.000	0.000	0.000	1.000	7.27 ± 0.07	0.000
SMUCE	fms(0.2)	Yes	0.000	0.000	0.000	0.000	1.000	8.02 ± 0.11	0.000
CL	fms(0.2)	Yes	0.000	0.000	0.000	0.000	1.000	7.29 ± 0.07	0.000
WBS sSIC	fms(0.2)	Yes	0.000	0.000	0.000	0.000	1.000	7.33 ± 0.07	0.000
FDR($\alpha = 0.05$)	fms(0.2)	Yes	0.000	0.000	0.000	0.000	1.000	7.44 ± 0.07	0.000
robseg(Huber)	fms(0.2)	Yes	0.000	0.000	0.000	0.000	1.000	7.51 ± 0.08	0.000
robseg(biweight)	fms(0.2)	Yes	0.000	0.000	**0.956**	0.043	0.001	1.64 ± 0.06	1.000

Table 4.2: Frequencies of $\widehat{N} - N$ and $\widehat{R}_n(\cdot)$ of ES and its competitors for `fms` ($\sigma = 0.2$) signal with 5 outliers over 1000 simulated sample paths. Contribution denotes the frequency of each competitor being selected as ES. Bold: highest empirical frequency of $\widehat{N} - N = 0$. The uncertainty is obtained by computing $2\widehat{\sigma}/\sqrt{n_r}$, where $\widehat{\sigma}^2$ is the empirical variance and n_r is the number of replications.

We can observe from Table 4.2 that in such a scenario PELT, SMUCE, CL, WBS sSIC, FDR and robseg (Huber) are all not robust with respect to the outliers and they all overestimate the number of changepoints due to fitting the outliers. The CBS and cumSeg still systematically underestimate the number of changepoints. It is not that surprising robseg (biweight) proposed in Fearnhead and Rigaill (2019) is quite robust in this scenario since it was designed to handle such an issue. It shows a very high frequency 0.956 to recover the correct number of changepoints. Moreover, from the quantity of empirical risk $\widehat{R}_n(\cdot)$, it turns out robseg (biweight) outperforms all the other candidates significantly which also indicates an excellent performance of localising the changepoints as well as estimating the value of $\boldsymbol{\gamma}^\star$ on each segment. Our selection procedure automatically gives the preference to robseg (biweight) in this case with frequency 1.000 which confirms the stability of our selection rule practically.

4.4.3 From Gaussian to Poisson and exponential models

As we have mentioned in Section 4.1, there are not too many work in the statistical literature addressing changepoint detection for Poisson and exponential models and establishing

a theoretical guarantee for the proposed estimator. The CL method proposed by Cleynen and Lebarbier (2014, 2017)) performs a model selection procedure based on the partitions given by Rigaill (2015) and they have proved the resulting estimator satisfies some oracle inequality. Implementing their procedure, we find the R package `Segmentor3IsBack` tackling both of Poisson and exponential models. Another approach is given by Frick et al. (2013) with the R package `StepR` where algorithm is only available for Poisson segmentation.

Recall that in Section 4.2.2, one feature of our selection procedure is enlarging the (possibly random) collection $\widehat{\mathbf{\Gamma}}$ but keeping $\mathcal{M}$ unchanged, the risk bound for the selected estimator only decreases (or at least keeps unchanged) over the larger collection. Therefore, for Poisson and exponential models, besides CL and SMUCE (if available), we would like to recruit some reasonable estimators into our candidates set $\widehat{\mathbf{\Gamma}}$. Although these estimators do not exist in the literature and no quantitative or qualitative analysis for them, once they are selected as ES by our selection procedure, the theoretical guarantee we built in Section 4.2.2 indicates that, up to a constant, they perform better than the state-of-art ones (CL and SMUCE).

One natural idea is to borrow the estimators for Gaussian model which is the case intensively studied. Inspired by Brown et al. (2010) where they implemented a mean-matching variance stabilizing transformation (MM-VST for short) to turn the problem of regression in exponential families into a standard homoscedastic Gaussian regression problem, we can perform a similar technique to the observations $\boldsymbol{Y}$. For more details of MM-VST, we refer to Section 2 of Brown et al. (2010). Let us remark that while implementing MM-VST, we need to choose the value of m which corresponds to the number of data-points binned for transformation. Although it turns out that for regression problem, this m needs to be suitably chosen (see Section 4 of Brown et al. (2010)), we do not want this pre-process step presumes any information of the segmentation as we are in the context of changepoint detection. Therefore, we simply take $m = 1$ in their transformation procedure and implement the formula $Y_i' = 2\sqrt{Y_i + 1/4}$ for Poisson model and $Y_i' = \log(2Y_i)$ for exponential model to derive new sequences of observations $\boldsymbol{Y}' = (Y_1', \ldots, Y_n')$. We then apply the algorithms introduced in the last section to $\boldsymbol{Y}'$ to get the locations of changepoints. Based on these locations, we associate ρ-estimators introduced in Chapter 2 to the estimated values of $\boldsymbol{\gamma}^\star$ on each segment to improve the performance. Recall that as we have seen in Section 2.5, under some suitable conditions and when the model is exact, ρ-estimator recovers the accurate result given by MLE. Moreover, it possesses more robustness compared to MLE when there is a model misspecification

and/or data contamination. To conclude, the candidates set for Poisson model is given by

$$\widehat{\mathbf{\Gamma}} = \Big\{\text{SMUCE}, \text{CL}, \text{CBS}^{\text{t}} + \rho, \text{cumSeg}^{\text{t}} + \rho, \text{PELT}^{\text{t}} + \rho, \text{WBS sSIC}^{\text{t}} + \rho, \\ \text{FDR}^{\text{t}}(\alpha = 0.05) + \rho, \text{robseg(Huber)}^{\text{t}} + \rho, \text{robseg(biweight)}^{\text{t}} + \rho \Big\}, \tag{4.4.1}$$

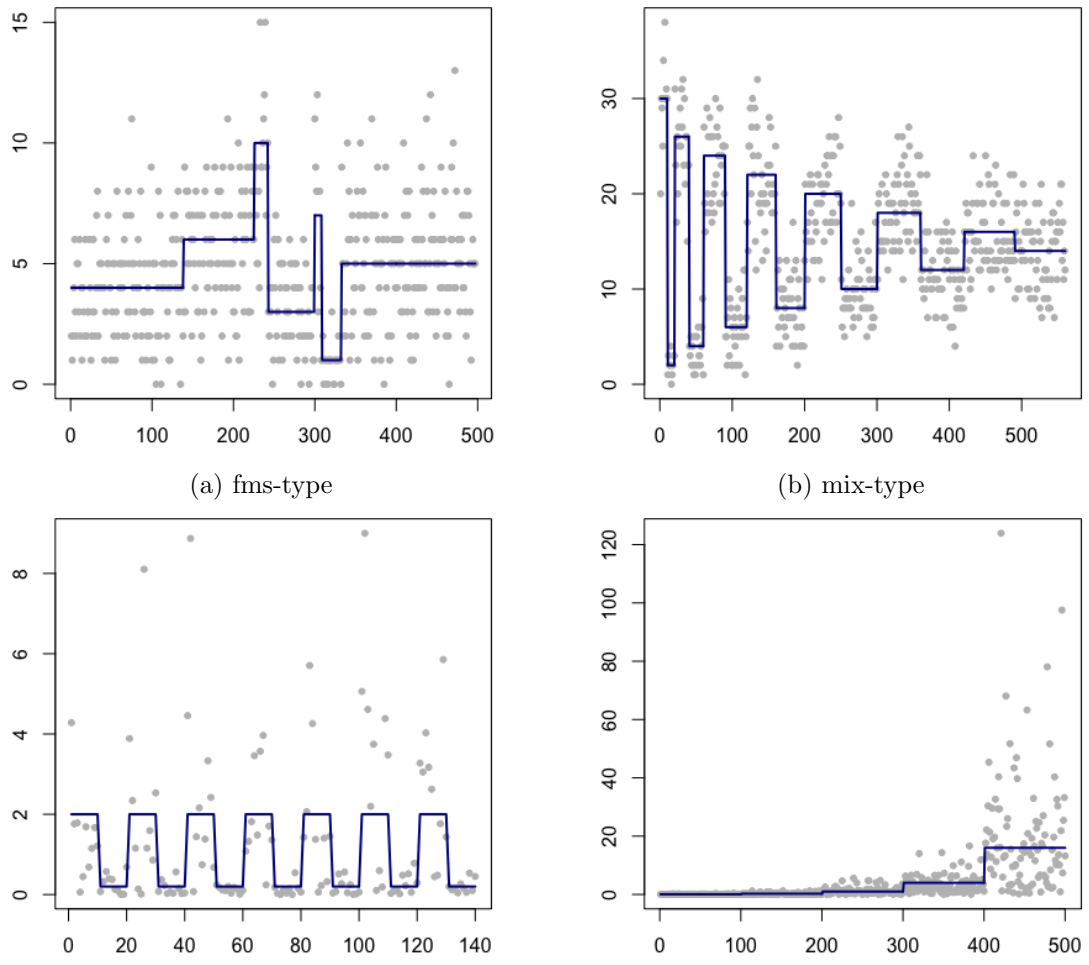
where the character "t" indicates the procedure is implemented on the transformed data. For exponential model, $\widehat{\mathbf{\Gamma}}$ is constructed the same as (4.4.1) except we change SMUCE to $\text{SMUCE}^{\text{t}} + \rho$ since it is no longer available.

To investigate the performance of ES and the candidates in $\widehat{\mathbf{\Gamma}}$, we mimic the test signals `fms` and `mix` for Poisson model and `teeth10` and `stairs10` for exponential model. We also study the scenario when outliers present in the observations for the mimic signals `fms` (Poisson) and `teeth10` (exponential). We shall describe the specific settings of these signals as well as how we add outliers in Section 4.7. Figure 4.4 exhibits the four underlying signals together with one profile of the simulated data for each signal.

			$\widehat{N} - N$						
Method	Signal	Outlier	≤ -2	-1	0	1	≥ 2	$\widehat{R}_n(\cdot)$	Contribution
ES	fms-type	No	0.002	0.050	**0.878**	0.062	0.008	2.51 ± 0.09	-
SMUCE	fms-type	No	0.288	0.528	0.184	0.000	0.000	6.21 ± 0.19	0.184
CL	fms-type	No	0.000	0.046	**0.854**	0.082	0.018	2.54 ± 0.10	0.725
$\text{CBS}^{\text{t}} + \rho$	fms-type	No	0.030	0.254	0.546	0.131	0.039	5.34 ± 0.11	0.000
$\text{cumSeg}^{\text{t}} + \rho$	fms-type	No	0.424	0.374	0.193	0.009	0.000	7.26 ± 0.20	0.001
$\text{PELT}^{\text{t}} + \rho$	fms-type	No	0.003	0.054	**0.867**	0.062	0.014	2.56 ± 0.10	0.015
$\text{WBS sSIC}^{\text{t}} + \rho$	fms-type	No	0.010	0.132	0.781	0.051	0.026	3.08 ± 0.12	0.013
$\text{FDR}^{\text{t}}(\alpha = 0.05) + \rho$	fms-type	No	0.288	0.528	0.184	0.000	0.000	5.97 ± 0.18	0.000
$\text{robseg(Huber)}^{\text{t}} + \rho$	fms-type	No	0.001	0.035	**0.800**	0.130	0.034	2.68 ± 0.10	0.032
$\text{robseg(biweight)}^{\text{t}} + \rho$	fms-type	No	0.005	0.073	**0.867**	0.048	0.007	2.63 ± 0.10	0.030
ES	fms-type	Yes	0.001	0.092	**0.825**	0.070	0.012	3.78 ± 0.11	-
SMUCE	fms-type	Yes	0.000	0.000	0.000	0.000	1.000	12.76 ± 0.21	0.000
CL	fms-type	Yes	0.000	0.000	0.000	0.000	1.000	8.58 ± 0.11	0.000
$\text{CBS}^{\text{t}} + \rho$	fms-type	Yes	0.521	0.354	0.086	0.035	0.004	11.98 ± 0.36	0.000
$\text{cumSeg}^{\text{t}} + \rho$	fms-type	Yes	0.795	0.164	0.038	0.003	0.000	12.01 ± 0.28	0.009
$\text{PELT}^{\text{t}} + \rho$	fms-type	Yes	0.000	0.000	0.000	0.000	1.000	8.45 ± 0.10	0.000
$\text{WBS sSIC}^{\text{t}} + \rho$	fms-type	Yes	0.000	0.000	0.000	0.000	1.000	8.82 ± 0.12	0.000
$\text{FDR}^{\text{t}}(\alpha = 0.05) + \rho$	fms-type	Yes	0.000	0.000	0.000	0.000	1.000	11.35 ± 0.18	0.001
$\text{robseg(Huber)}^{\text{t}} + \rho$	fms-type	Yes	0.000	0.008	0.048	0.062	0.882	6.13 ± 0.13	0.053
$\text{robseg(biweight)}^{\text{t}} + \rho$	fms-type	Yes	0.000	0.092	**0.839**	0.066	0.003	3.74 ± 0.11	0.937
ES	mix-type	No	0.005	0.371	**0.523**	0.091	0.010	3.98 ± 0.09	-
SMUCE	mix-type	No	0.128	0.828	0.044	0.000	0.000	4.67 ± 0.13	0.339
CL	mix-type	No	0.014	0.439	0.466	0.071	0.010	3.99 ± 0.09	0.481
$\text{CBS}^{\text{t}} + \rho$	mix-type	No	0.034	0.448	0.358	0.122	0.038	13.39 ± 0.11	0.000
$\text{cumSeg}^{\text{t}} + \rho$	mix-type	No	0.990	0.010	0.000	0.000	0.000	31.18 ± 0.45	0.000
$\text{PELT}^{\text{t}} + \rho$	mix-type	No	0.010	0.443	0.466	0.071	0.010	4.03 ± 0.09	0.027
$\text{WBS sSIC}^{\text{t}} + \rho$	mix-type	No	0.018	0.509	0.402	0.056	0.015	4.03 ± 0.09	0.013
$\text{FDR}^{\text{t}}(\alpha = 0.05) + \rho$	mix-type	No	0.128	0.828	0.044	0.000	0.000	4.67 ± 0.12	0.000
$\text{robseg(Huber)}^{\text{t}} + \rho$	mix-type	No	0.003	0.293	**0.530**	0.149	0.025	4.15 ± 0.09	0.099
$\text{robseg(biweight)}^{\text{t}} + \rho$	mix-type	No	0.014	0.486	0.458	0.040	0.002	4.06 ± 0.09	0.041

Table 4.3: Frequencies of $\widehat{N} - N$ and $\widehat{R}_n(\cdot)$ of ES and its competitors for Poisson model over 1000 simulated sample paths. Contribution denotes the frequency of each competitor being selected as ES. Bold: highest empirical frequency of $\widehat{N} - N = 0$ and those with frequencies within 10% off the highest. The uncertainty is obtained by computing $2\widehat{\sigma}/\sqrt{n_r}$, where $\widehat{\sigma}^2$ is the empirical variance and n_r is the number of replications.

The results of Poisson model are shown in Table 4.3. Let us first comment the two



(a) fms-type (b) mix-type

(c) teeth-type (d) stairs-type

Figure 4.4: (A) and (B): the test signals of the form $\exp(\boldsymbol{\gamma}^\star)$ (solid line) and simulated data (dots) for Poisson model. (C) and (D): the test signals of the form of $1/\boldsymbol{\gamma}^\star$ (solid line) and simulated data (dots) for exponential model.

existing estimators in the literature, namely SMUCE and CL. In both of the scenarios with or without outliers, the performance of SMUCE is quite poor at least under these two test signals. When no outlier presents in the observations, SMUCE has a tendency to underestimate the number of changepoints for both of the two signals `fms-type` and `mix-type`. When there are outliers, SMUCE is sensitive to them therefore overestimates the number of changepoints. The CL performs much better than SMUCE in the scenario that no outlier presents in the observations but it is also not robust with respect to the outliers. When no outlier presents, our estimator ES slightly improves the performance of CL on detecting changes under both of the two signals. When the outliers present, ES obviously outperforms CL as a consequence of enjoying the excellent performance given

by robseg (biweight)$^{\mathrm{t}}$. Interestingly, we find that when there is no outlier presenting in the observations, the combinations PELT$^{\mathrm{t}}+\rho$ and robseg$^{\mathrm{t}}+\rho$ are perhaps nice choices at least under these two signals.

The results for exponential model are shown in Table 4.4. Under the `teeth-type` signal without an outlier, the ES obviously outperforms any single candidate by selecting mainly from CL and robseg (biweight)$^{\mathrm{t}}+\rho$. When there are outliers, robseg (biweight)$^{\mathrm{t}}+\rho$ is the best one among all and we observe that ES improves the frequency to select robseg (biweight)$^{\mathrm{t}}+\rho$ as the final estimator so that finally ES achieves a competitive performance compared to robseg (biweight)$^{\mathrm{t}}+\rho$ and significantly outperforms the existing estimator CL. For `stairs-type` signal, CL performs quite nice but ES still slightly improves it by enjoying the contribution from other candidates in $\widehat{\mathbf{\Gamma}}$.

			$\widehat{N}-N$						
Method	Signal	Outlier	≤ -2	-1	0	1	≥ 2	$\widehat{R}_n(\cdot)$	Contribution
ES	teeth-type	No	0.327	0.077	**0.468**	0.106	0.022	7.69 ± 0.25	-
CL	teeth-type	No	0.381	0.055	0.411	0.116	0.037	9.27 ± 0.38	0.766
SMUCE$^{\mathrm{t}}+\rho$	teeth-type	No	0.998	0.002	0.000	0.000	0.000	20.29 ± 0.14	0.000
CBS$^{\mathrm{t}}+\rho$	teeth-type	No	1.000	0.000	0.000	0.000	0.000	22.52 ± 0.06	0.000
cumSeg$^{\mathrm{t}}+\rho$	teeth-type	No	1.000	0.000	0.000	0.000	0.000	22.56 ± 0.05	0.000
PELT$^{\mathrm{t}}+\rho$	teeth-type	No	0.134	0.068	0.241	0.191	0.366	9.14 ± 0.19	0.015
WBS sSIC$^{\mathrm{t}}+\rho$	teeth-type	No	0.829	0.022	0.058	0.036	0.055	18.62 ± 0.37	0.007
FDR$^{\mathrm{t}}(\alpha=0.05)+\rho$	teeth-type	No	0.998	0.002	0.000	0.000	0.000	20.30 ± 0.14	0.000
robseg(Huber)$^{\mathrm{t}}+\rho$	teeth-type	No	0.076	0.096	0.263	0.227	0.338	8.42 ± 0.18	0.023
robseg(biweight)$^{\mathrm{t}}+\rho$	teeth-type	No	0.435	0.122	0.348	0.082	0.013	8.86 ± 0.21	0.189
ES	teeth-type	Yes	0.383	0.082	**0.303**	0.151	0.081	9.38 ± 0.25	-
CL	teeth-type	Yes	0.500	0.048	0.169	0.128	0.155	12.42 ± 0.42	0.534
SMUCE$^{\mathrm{t}}+\rho$	teeth-type	Yes	1.000	0.000	0.000	0.000	0.000	22.02 ± 0.14	0.000
CBS$^{\mathrm{t}}+\rho$	teeth-type	Yes	1.000	0.000	0.000	0.000	0.000	24.43 ± 0.07	0.001
cumSeg$^{\mathrm{t}}+\rho$	teeth-type	Yes	1.000	0.000	0.000	0.000	0.000	24.49 ± 0.06	0.000
PELT$^{\mathrm{t}}+\rho$	teeth-type	Yes	0.090	0.069	0.131	0.162	0.548	10.48 ± 0.18	0.023
WBS sSIC$^{\mathrm{t}}+\rho$	teeth-type	Yes	0.908	0.014	0.017	0.024	0.037	21.82 ± 0.32	0.008
FDR$^{\mathrm{t}}(\alpha=0.05)+\rho$	teeth-type	Yes	1.000	0.000	0.000	0.000	0.000	22.02 ± 0.14	0.001
robseg(Huber)$^{\mathrm{t}}+\rho$	teeth-type	Yes	0.090	0.074	0.202	0.222	0.412	9.37 ± 0.18	0.105
robseg(biweight)$^{\mathrm{t}}+\rho$	teeth-type	Yes	0.456	0.106	**0.316**	0.105	0.017	9.48 ± 0.20	0.328
ES	stairs-type	No	0.000	0.000	**0.923**	0.067	0.010	2.09 ± 0.08	-
CL	stairs-type	No	0.000	0.000	**0.907**	0.075	0.018	2.10 ± 0.08	0.977
SMUCE$^{\mathrm{t}}+\rho$	stairs-type	No	0.000	0.008	0.489	0.225	0.278	4.28 ± 0.19	0.003
CBS$^{\mathrm{t}}+\rho$	stairs-type	No	0.006	0.134	0.594	0.193	0.073	6.27 ± 0.31	0.000
cumSeg$^{\mathrm{t}}+\rho$	stairs-type	No	0.002	0.120	0.682	0.192	0.004	7.54 ± 0.32	0.000
PELT$^{\mathrm{t}}+\rho$	stairs-type	No	0.000	0.000	0.032	0.041	0.927	6.56 ± 0.18	0.000
WBS sSIC$^{\mathrm{t}}+\rho$	stairs-type	No	0.000	0.003	0.456	0.094	0.447	4.63 ± 0.17	0.001
FDR$^{\mathrm{t}}(\alpha=0.05)+\rho$	stairs-type	No	0.000	0.008	0.489	0.225	0.278	4.27 ± 0.19	0.002
robseg(Huber)$^{\mathrm{t}}+\rho$	stairs-type	No	0.000	0.000	0.207	0.144	0.649	4.84 ± 0.16	0.001
robseg(biweight)$^{\mathrm{t}}+\rho$	stairs-type	No	0.000	0.000	0.699	0.183	0.118	3.21 ± 0.12	0.016

Table 4.4: Frequencies of $\widehat{N}-N$ and $\widehat{R}_n(\cdot)$ of ES and its competitors for exponential model over 1000 simulated sample paths. Contribution denotes the frequency of each competitor being selected as ES. Bold: highest empirical frequency of $\widehat{N}-N=0$ and those with frequencies within 10% off the highest. The uncertainty is obtained by computing $2\widehat{\sigma}/\sqrt{n_r}$, where $\widehat{\sigma}^2$ is the empirical variance and n_r is the number of replications.

4.5 Real data examples

In this section, we apply our estimator selection procedure to two real datasets and investigate its performance. The first one is the observations of DNA copy numbers from biological research where Gaussian model is considered to detect changes. The second one is the British coal disasters dataset to which Poisson model is applied.

4.5.1 Detecting changes in DNA copy numbers

In normal human cells, it is well known that the number of DNA copies is two. As it has been revealed by much work in biological research (see Albertson and Pinkel (2003) and Redon et al. (2006) for example), the pathogenesis of some diseases including various cancers and mental retardation is often associated to chromosomal aberrations such as deletions, duplications and/or amplifications which finally result in the copy number of DNA from such regions differs from the normal number two. Including microarray and sequencing experiments, biologists have developed various techniques to measure DNA copy numbers of selected genes on some genome and they record their experimental results as a sequence of observations $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. The interest lies in finding abrupt changes in the means of the observations. To address this issue, we consider Gaussian model with an estimated variance.

In R package `jointseg` (Pierre-Jean et al. (2015)), they provide two real datasets GSE11976 and GSE29172 to resample from, where the truth of changepoints is already known. However, since we do not have the information of the associated value on each segment, it is impossible to compute the pseudo Hellinger distance between each estimator and the truth. Note that for both GSE11976 and GSE29172 datasets, we need to choose the tumour fraction when resampling from them. We consider the tumour fraction levels 0.79 and 1 for the dataset GSE11976 and the levels 0.7 and 1 for GSE29172 which turns out to be the situations where the size of each jump at the changepoint is relatively large as indicated in Figure 9 of Fearnhead and Rigaill (2019). Therefore, we can roughly evaluate the performance of each estimator by its frequency of correctly estimating the number of changepoints. Although our selection procedure can be applied in the scenario where small amount of outliers present in the observations, as we have seen in Section 4.4 some candidates in $\widehat{\boldsymbol{\Gamma}}$ are sensitive to the outliers. To avoid the phenomenon that an estimator systematically underestimates the number of changepoints but due to the sensitivity to outliers it accidentally gives a correct number of segments, we run a smooth procedure on the data before applying all the estimation procedures by implementing the function `smooth.CNA` from the famous R package `DNAcopy`. Moreover, since we have seen in the simulation study that the performance of CBS and cumSeg is quite poor, we remove these two estimators from our candidates set $\widehat{\boldsymbol{\Gamma}}$ for simplicity. For each dataset and each level of tumour fraction, we simulate 1000 profiles of length $n = 1000$ with 5 changepoints where

the length of each segment is at least 20. The results are shown in Table 4.5. As one can observe, among the state-of-art ones, robseg (biweight) is the best for correctly estimating the number of changepoints on this dataset. By running a data-driven procedure to select among the candidates set $\widehat{\mathbf{\Gamma}}$, our selected estimator ES shows a competitive performance in this situation as compared to the best one robseg (biweight).

			$\widehat{N}-N$							
Method	Dataset	Fraction	≤ -3	-2	-1	0	1	2	≥ 3	Contribution
ES	GSE11976	0.79	0.003	0.028	0.044	**0.771**	0.108	0.031	0.015	-
PELT	GSE11976	0.79	0.000	0.002	0.008	0.198	0.096	0.196	0.500	0.060
SMUCE	GSE11976	0.79	0.004	0.021	0.124	0.391	0.203	0.139	0.118	0.147
CL	GSE11976	0.79	0.011	0.066	0.053	0.550	0.117	0.118	0.085	0.393
WBS sSIC	GSE11976	0.79	0.005	0.031	0.066	0.508	0.066	0.174	0.150	0.100
FDR($\alpha = 0.05$)	GSE11976	0.79	0.000	0.005	0.011	0.096	0.056	0.126	0.706	0.020
robseg(Huber)	GSE11976	0.79	0.001	0.012	0.022	0.569	0.193	0.110	0.093	0.121
robseg(biweight)	GSE11976	0.79	0.002	0.046	0.045	**0.778**	0.102	0.019	0.008	0.159
ES	GSE11976	1.00	0.000	0.003	0.007	**0.790**	0.100	0.046	0.054	-
PELT	GSE11976	1.00	0.000	0.000	0.000	0.243	0.067	0.195	0.495	0.046
SMUCE	GSE11976	1.00	0.000	0.002	0.035	0.395	0.178	0.177	0.213	0.208
CL	GSE11976	1.00	0.001	0.011	0.011	0.604	0.098	0.170	0.105	0.357
WBS sSIC	GSE11976	1.00	0.000	0.003	0.008	0.536	0.060	0.225	0.168	0.075
FDR($\alpha = 0.05$)	GSE11976	1.00	0.000	0.000	0.004	0.138	0.059	0.126	0.673	0.018
robseg(Huber)	GSE11976	1.00	0.000	0.002	0.004	0.559	0.163	0.126	0.146	0.155
robseg(biweight)	GSE11976	1.00	0.000	0.010	0.006	**0.794**	0.101	0.043	0.046	0.141
ES	GSE29172	0.70	0.014	0.136	0.133	**0.596**	0.088	0.028	0.005	-
PELT	GSE29172	0.70	0.003	0.027	0.054	0.210	0.139	0.181	0.386	0.089
SMUCE	GSE29172	0.70	0.016	0.112	0.307	0.247	0.176	0.087	0.055	0.099
CL	GSE29172	0.70	0.035	0.159	0.155	0.305	0.129	0.126	0.091	0.302
WBS sSIC	GSE29172	0.70	0.022	0.105	0.155	0.290	0.113	0.173	0.142	0.046
FDR($\alpha = 0.05$)	GSE29172	0.70	0.003	0.024	0.075	0.133	0.112	0.133	0.520	0.032
robseg(Huber)	GSE29172	0.70	0.007	0.068	0.087	0.533	0.163	0.092	0.050	0.224
robseg(biweight)	GSE29172	0.70	0.018	0.168	0.153	**0.597**	0.052	0.012	0.000	0.208
ES	GSE29172	1.00	0.000	0.005	0.003	**0.828**	0.093	0.051	0.020	-
PELT	GSE29172	1.00	0.000	0.001	0.001	0.233	0.070	0.251	0.444	0.046
SMUCE	GSE29172	1.00	0.000	0.004	0.044	0.416	0.193	0.199	0.144	0.185
CL	GSE29172	1.00	0.001	0.009	0.006	0.684	0.077	0.163	0.060	0.427
WBS sSIC	GSE29172	1.00	0.000	0.006	0.009	0.576	0.051	0.230	0.128	0.070
FDR($\alpha = 0.05$)	GSE29172	1.00	0.000	0.001	0.002	0.119	0.063	0.133	0.682	0.018
robseg(Huber)	GSE29172	1.00	0.000	0.001	0.001	0.594	0.145	0.158	0.101	0.120
robseg(biweight)	GSE29172	1.00	0.000	0.007	0.006	**0.833**	0.098	0.043	0.013	0.134

Table 4.5: Frequencies of $\widehat{N} - N$ of ES and its competitors for DNA copy numbers data. Contribution denotes the frequency of each competitor being selected as ES. Bold: highest empirical frequency of $\widehat{N} - N = 0$ and those with frequencies within 10% off the highest.

4.5.2 British coal disasters dataset

To investigate the performance of ES for Poisson model in practice, we apply our procedure to British coal disasters dataset. This dataset is quite well-known in the context of Poisson segmentation see Green (1995), Yang and Kuo (2001), Fearnhead (2006) and Lloyd et al. (2015) for example. We choose this dataset mainly because of two reasons. First, the changepoints have been studied by many different methods which makes it easier to understand our result. Besides, the sequence has a general tendency to decrease with the progress over time which can be correlated to implementing safety regulation in

the history. Though pretty rough, we have some evidence to evaluate the changepoint detection procedures on this dataset.

The data at hand include the number of each year coal disasters in UK during the period from March 15th, 1851 to March 22nd, 1962 with length $n = 112$. In this situation, Poisson model is considered based on the same candidates set (4.4.1) as described in Section 4.4.3. We conclude the results of different estimators as follows. Concerning to the changepoints, there are in total three suggestions:

(1) 1 changepoint at the year 1891: $\text{cumSeg}^{\text{t}}+\rho$, $\text{PELT}^{\text{t}}+\rho$, WBS $\text{sSIC}^{\text{t}}+\rho$, $\text{FDR}^{\text{t}}(\alpha = 0.05)+\rho$ and $\text{robseg(biweight)}^{\text{t}}+\rho$;

(2) 2 changepoints at the year 1891 and 1947: SMUCE and CL;

(3) 3 changepoints at the year 1891, 1929 and 1942: $\text{robseg(Huber)}^{\text{t}}+\rho$.

Our selection procedure finally choose SMUCE as ES, i.e. we support the suggestion with two changepoints at the year 1891 and 1947. The dataset as well as the result of ES (SMUCE) is plotted in Figure 4.5.

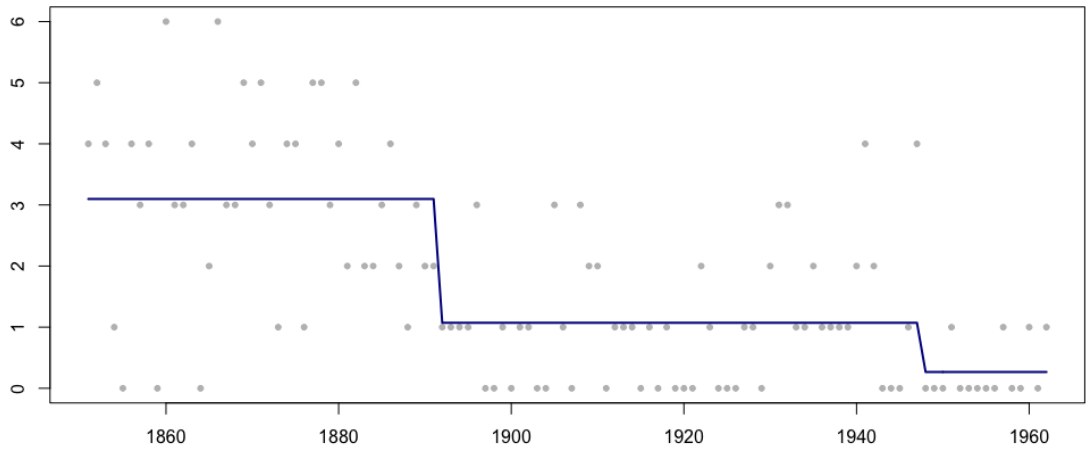


Figure 4.5: Coal mining disasters data (dots) and ES estimator (solid line).

Now we comment our result by comparing it with the existing ones in the literature. In Green (1995), they used the coal mining disasters data recorded per day and proposed a reversible jump MCMC approach to detect changepoints as well as estimating intensity function. According to the Figure 2 in the same paper, the model with two changepoints has the highest posterior probability. Moreover, according to their Figure 3, in the two changepoints scenario, the posterior mode is approximately 14,000 days for the first changepoint and 35,000 days for the second one. This is very close to our result since counting from March 15th, 1851, 14,000 days is between the year 1889 and 1890 and 35,000 days is the time between the year 1946 and 1947. Later, a Bayesian binary

segmentation procedure was proposed by Yang and Kuo (2001) to locate changepoints for Poisson process. Based on two different tests they adopted, their procedure obtained two different sets of changepoints (one changepoint for applying Bayes factor criterion and two for applying BIC approximation criterion) where the locations of changepoints for these two models are quite similar to the results (1) and (2) mentioned in the last paragraph. On the other hand, as it was pointed out in Lloyd et al. (2015), UK parliament passed several acts to improve the safety of mine works including the Coal Mines Regulation Acts of 1872 and 1887 and a further one in 1954 with mines and quarries acts. In general, it is reasonable to have a non-increasing expectation of the number of disasters after the year releasing these regulations. As it is shown in Figure 4.5, the model with two changepoints meets the releasing regulation years 1887 and 1954. Considering the best fit with the time of released regulations and the results given in the literatures, we believe the two changepoints model for this dataset is the most reasonable one to the truth.

4.6 Proofs of main and auxiliary results

Recall that $(\mathscr{X}, \mathcal{X}) = (\mathscr{W} \times \mathscr{Y}, \mathcal{W} \otimes \mathcal{Y})$ and $\mathscr{P}^f$ the set of all product probabilities on $(\mathscr{X}^n, \mathcal{X}^{\otimes n})$. For all $i \in \{1, \ldots, n\}$, we denote the true distribution of $X_i = (W_i, Y_i)$ by $P_i^\star$ and denote the true joint distribution of $\boldsymbol{X} = (X_1, \ldots, X_n)$ by $\mathbf{P}^\star = \otimes_{i=1}^n P_i^\star \in \mathscr{P}^f$. We denote $\mathbf{P}_{\boldsymbol{\gamma}} = \otimes_{i=1}^n P_{i,\boldsymbol{\gamma}}$ the joint 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 $R_{\boldsymbol{\gamma}(W_i)} \in \overline{\mathscr{Q}}$ for all $i \in \{1, \ldots, n\}$. Under such a notation setting, we have $P_i^\star = Q_i^\star \cdot P_{W_i}$, $P_{i,\boldsymbol{\gamma}} = R_{\boldsymbol{\gamma}} \cdot P_{W_i}$ as well as the following equality

$$h^2(P_i^\star, P_{i,\boldsymbol{\gamma}}) = \int_{\mathscr{W}} h^2(Q_i^\star(w), R_{\boldsymbol{\gamma}(w)}) dP_{W_i}(w). \tag{4.6.1}$$

As an immediate consequence of (4.6.1) and (1.4.1), for any $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}$,

$$\begin{aligned} \mathbf{h}^2(\mathbf{Q}^\star, \mathbf{R}_{\boldsymbol{\gamma}}) &= \sum_{i=1}^n \int_{\mathscr{W}} h^2(Q_i^\star(w), R_{\boldsymbol{\gamma}(w)}) dP_{W_i}(w) \\ &= \sum_{i=1}^n h^2(P_i^\star, P_{i,\boldsymbol{\gamma}}) = \mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\boldsymbol{\gamma}}). \end{aligned} \tag{4.6.2}$$

For each $m \in \mathcal{M}$, we define the set of probabilities $\mathscr{P}_m = \{\mathbf{P}_{\boldsymbol{\gamma}},\ \boldsymbol{\gamma} \in \boldsymbol{\Gamma}_m\}$ and $\mathscr{P} = \{\mathbf{P}_{\boldsymbol{\gamma}},\ \boldsymbol{\gamma} \in \boldsymbol{\Gamma}\}$ with $\boldsymbol{\Gamma} = \cup_{m \in \mathcal{M}} \boldsymbol{\Gamma}_m$. For any $y > 0$, $\mathbf{P}^\star \in \mathscr{P}^f$, $\mathscr{P}_{m_1}, \mathscr{P}_{m_2}$ with $m_1, m_2 \in \mathcal{M}$, we define the set

$$\begin{aligned} &\mathscr{B}^{\mathscr{P}_{m_1} \times \mathscr{P}_{m_2}}(\mathbf{P}^\star, y) \\ &= \left\{(\mathbf{P}_{\boldsymbol{\gamma}_1}, \mathbf{P}_{\boldsymbol{\gamma}_2}) \middle|\ \mathbf{P}_{\boldsymbol{\gamma}_1} \in \mathscr{P}_{m_1}, \mathbf{P}_{\boldsymbol{\gamma}_2} \in \mathscr{P}_{m_2}, \mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\boldsymbol{\gamma}_1}) + \mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\boldsymbol{\gamma}_2}) < y^2\right\} \end{aligned}$$

and for any $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2 \in \boldsymbol{\Gamma}$, we set

$$\mathbf{Z}(\mathbf{X}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) = \mathbf{T}(\mathbf{X}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) - \mathbb{E}\left[\mathbf{T}(\mathbf{X}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)\right].$$

We then introduce bellow Proposition 45 of Baraud et al. (2017) which is an extensional version of Talagrand's Theorem on the supremum of empirical processes proved in Massart (2007).

Proposition 4.6.1. *Let T be some finite or countable set, U_1,...,U_n be independent centered random vectors with values in $\mathbb{R}^T$ and let*

$$Z = \sup_{t \in T} \Big| \sum_{i=1}^{n} U_{i,t} \Big|.$$

If for some positive numbers b and v,

$$\max_{i=1,\dots,n} |U_{i,t}| \leq b \quad \text{and} \quad \sum_{i=1}^{n} \mathbb{E}\left[U_{i,t}^2\right] \leq v^2 \quad \text{for all } t \in T,$$

then, for all positive numbers c and x,

$$\mathbb{P}\left[Z \leq (1+c)\mathbb{E}(Z) + (8b)^{-1}cv^2 + 2(1+8c^{-1})bx\right] \geq 1 - e^{-x}.$$

4.6.1 Elementary results and proofs

Before showing the main theorem, we present two preliminary results and their proofs in this section.

Lemma 4.6.1. *Let $m_1, m_2 \in \mathcal{M}$ be two partitions on $\mathscr{W}$. The class of functions*

$$\mathscr{F}(m_1, m_2) = \left\{ \frac{r_{\boldsymbol{\gamma}_2}}{r_{\boldsymbol{\gamma}_1}} : (w, y) \mapsto \frac{r_{\boldsymbol{\gamma}_2(w)}(y)}{r_{\boldsymbol{\gamma}_1(w)}(y)},\ \boldsymbol{\gamma}_1 \in \boldsymbol{\Gamma}_{m_1},\ \boldsymbol{\gamma}_2 \in \boldsymbol{\Gamma}_{m_2} \right\}$$

on $\mathscr{X} = \mathscr{W} \times \mathscr{Y}$ is VC-subgraph with dimension not larger than $2|m_1 \vee m_2| + 1$.

Proof. Recall that $r_{\boldsymbol{\gamma}} : (w, y) \mapsto \exp\left[u(\boldsymbol{\gamma}(w))S(y) - B(\boldsymbol{\gamma}(w))\right]$ according to (2.3.12). For any $\boldsymbol{\gamma}_1 \in \boldsymbol{\Gamma}_{m_1}$ and $\boldsymbol{\gamma}_2 \in \boldsymbol{\Gamma}_{m_2}$, we define function $g_{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2}$ on $\mathscr{W} \times \mathscr{Y}$ as

$$g_{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2}(w, y) = S(y)\left[u(\boldsymbol{\gamma}_2(w)) - u(\boldsymbol{\gamma}_1(w))\right] - \left[B(\boldsymbol{\gamma}_2(w)) - B(\boldsymbol{\gamma}_1(w))\right]$$

and define $\mathscr{G}(m_1, m_2)$ the class of functions as

$$\mathscr{G}(m_1, m_2) = \left\{ g_{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2} |\ \boldsymbol{\gamma}_1 \in \boldsymbol{\Gamma}_{m_1}, \boldsymbol{\gamma}_2 \in \boldsymbol{\Gamma}_{m_2} \right\}.$$

With the fact that $\mathscr{F}(m_1, m_2) = \{e^g,\ g \in \mathscr{G}(m_1, m_2)\}$ and the exponential function is monotone on $\mathbb{R}$, by Proposition 1.5.2 (Proposition 42 of Baraud et al. (2017)), it is enough to prove the conclusion holds for the class $\mathscr{G}(m_1, m_2)$.

Let $K = |m_1 \vee m_2|$ be the number of segments given by the refined partition $m_1 \vee m_2$ and $\mathcal{I}_1, \ldots, \mathcal{I}_K$ the resulted segments on $\mathscr{W}$. For any $\boldsymbol{\gamma}_1 \in \boldsymbol{\Gamma}_{m_1}$, we can rewrite it as

$$\boldsymbol{\gamma}_1(w) = \sum_{k=1}^{K} a_k \mathbb{1}_{\mathcal{I}_k}(w), \quad \text{where } (a_1, \ldots, a_K) \in J^K$$

and any $\boldsymbol{\gamma}_2 \in \boldsymbol{\Gamma}_{m_2}$,

$$\boldsymbol{\gamma}_2(w) = \sum_{k=1}^{K} b_k \mathbb{1}_{\mathcal{I}_k}(w), \quad \text{where } (b_1, \ldots, b_K) \in J^K.$$

As an immediate consequence, for any $g_{\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2} \in \mathscr{G}(m_1, m_2)$, it can be rewritten as

$$g_{\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2}(w, y) = \sum_{k=1}^{K} \left[u(b_k) - u(a_k)\right] \mathbb{1}_{\mathcal{I}_k}(w) S(y) - \sum_{k=1}^{K} \left[B(b_k) - B(a_k)\right] \mathbb{1}_{\mathcal{I}_k}(w).$$

Therefore, $\mathscr{G}(m_1, m_2)$ is contained in a $2K$-dimensional vector space spanned by

$$\left\{\mathbb{1}_{\mathcal{I}_k}(w), S(y)\mathbb{1}_{\mathcal{I}_k}(w),\ k = 1, \ldots, K\right\}.$$

By Proposition 1.5.1 (Lemma 2.6.15 of van der Vaart and Wellner (1996)), we conclude $\mathscr{G}(m_1, m_2)$ is VC-subgraph on $\mathscr{X} = \mathscr{W} \times \mathscr{Y}$ with dimension not larger than $2K + 1$. □

Proposition 4.6.2. *Let $m_1, m_2 \in \mathcal{M}$ be two partitions on $\mathscr{W}$. Under Assumption 4.2.2, for any $\mathbf{P}^\star \in \mathscr{P}^f$, $\eta \geq 1$ and any $y > 0$ satisfying*

$$y^2 \geq \eta \left[D_n(m_1) + D_n(m_2)\right],$$

we have

$$\mathbb{E}\left[\sup_{(\mathbf{P}_{\boldsymbol{\gamma}_1}, \mathbf{P}_{\boldsymbol{\gamma}_2}) \in \mathscr{B}^{\mathscr{P}_{m_1} \times \mathscr{P}_{m_2}}(\mathbf{P}^\star, y)} |\mathbf{Z}(\mathbf{X}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)|\right] \leq \left[9.77\sqrt{\frac{2\alpha + 1/2}{\eta}} + \frac{90(2\alpha + 1/2)}{\eta}\right] y^2.$$

Proof. We set $\boldsymbol{\mu} = \otimes_{i=1}^n \mu_i$ with $\mu_i = P_{W_i} \otimes \nu$. For any $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}$, we denote $\mathbf{r}_{\boldsymbol{\gamma}}$ a density (with respect to $\boldsymbol{\mu}$) on $\mathscr{X}^n = (\mathscr{W} \times \mathscr{Y})^n$ as

$$\mathbf{r}_{\boldsymbol{\gamma}}(x_1, \ldots, x_n) = r_{\boldsymbol{\gamma}}(x_1) \cdots r_{\boldsymbol{\gamma}}(x_n), \quad \text{for all } (x_1, \ldots, x_n) \in \mathscr{X}^n$$

so that for any $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}$, we have $\mathbf{P}_{\boldsymbol{\gamma}} = \mathbf{r}_{\boldsymbol{\gamma}} \cdot \boldsymbol{\mu}$. For any $y > 0$, we define $\mathscr{F}_y(m_1, m_2)$ the class of functions on $\mathscr{X}$ as

$$\left\{\psi\left(\sqrt{\frac{r_{\boldsymbol{\gamma}_2}}{r_{\boldsymbol{\gamma}_1}}}\right) \middle|\ \boldsymbol{\gamma}_1 \in \boldsymbol{\Gamma}_{m_1}, \boldsymbol{\gamma}_2 \in \boldsymbol{\Gamma}_{m_2}, \mathbf{h}^2(\mathbf{P}^\star, \mathbf{r}_{\boldsymbol{\gamma}_1} \cdot \boldsymbol{\mu}) + \mathbf{h}^2(\mathbf{P}^\star, \mathbf{r}_{\boldsymbol{\gamma}_2} \cdot \boldsymbol{\mu}) < y^2\right\}.$$

Since $\mathscr{F}_y(m_1, m_2)$ is a subset of the collection

$$\left\{\psi\left(\sqrt{\frac{r_{\boldsymbol{\gamma}_2}}{r_{\boldsymbol{\gamma}_1}}}\right) \middle|\ \boldsymbol{\gamma}_1 \in \boldsymbol{\Gamma}_{m_1}, \boldsymbol{\gamma}_2 \in \boldsymbol{\Gamma}_{m_2}\right\}$$

and the function ψ is monotone, it follows from Lemma 4.6.1 and Proposition 1.5.2 (Proposition 42-(ii) of Baraud et al. (2017)) that $\mathscr{F}_y(m_1, m_2)$ is VC-subgraph on $\mathscr{X}$ with dimension not larger than $\overline{V} = 2|m_1 \vee m_2| + 1$. Besides, by Proposition 3 of Baraud and Birgé (2018), our choice of the function ψ satisfies their Assumption 2 and more precisely (11) in their paper with $a_2^2 = 3\sqrt{2}$. Proposition 3 of Baraud and Birgé (2018) together with the definition of $\mathscr{F}_y$ implies that for any $y > 0$,

$$\sup_{f \in \mathscr{F}_y(m_1,m_2)} n^{-1} \sum_{i=1}^{n} \mathbb{E}\left[f^2(X_i)\right] \leq \frac{a_2^2 y^2}{n}. \tag{4.6.3}$$

Moreover, since the function ψ takes values in $[-1, 1]$, we derive from (4.6.3) that

$$\sup_{f \in \mathscr{F}_y(m_1,m_2)} n^{-1} \sum_{i=1}^{n} \mathbb{E}\left[f^2(X_i)\right] \leq \left(\frac{a_2^2 y^2}{n}\right) \wedge 1 \leq 1.$$

To bound the expectation of the supremum of an empirical process over a VC-subgraph class, we apply Theorem 2.6.1 in Chapter 2 to $\mathscr{F}_y(m_1, m_2)$ and obtain

$$\begin{aligned}
&\mathbb{E}\left[\sup_{(\mathbf{P}_{\boldsymbol{\gamma}_1},\mathbf{P}_{\boldsymbol{\gamma}_2}) \in \mathscr{B}^{\mathscr{P}_{m_1} \times \mathscr{P}_{m_2}}(\mathbf{P}^\star, y)} |\mathbf{Z}(\mathbf{X}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)|\right] \\
=&\mathbb{E}\left[\sup_{(\mathbf{P}_{\boldsymbol{\gamma}_1},\mathbf{P}_{\boldsymbol{\gamma}_2}) \in \mathscr{B}^{\mathscr{P}_{m_1} \times \mathscr{P}_{m_2}}(\mathbf{P}^\star, y)} |\mathbf{T}(\mathbf{X}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) - \mathbb{E}\left[\mathbf{T}(\mathbf{X}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)\right]|\right] \\
=&\mathbb{E}\left[\sup_{(\mathbf{P}_{\boldsymbol{\gamma}_1},\mathbf{P}_{\boldsymbol{\gamma}_2}) \in \mathscr{B}^{\mathscr{P}_{m_1} \times \mathscr{P}_{m_2}}(\mathbf{P}^\star, y)} \left|\sum_{i=1}^{n} \psi\left(\sqrt{\frac{r_{\boldsymbol{\gamma}_2(W_i)}(Y_i)}{r_{\boldsymbol{\gamma}_1(W_i)}(Y_i)}}\right) - \mathbb{E}\left[\sum_{i=1}^{n} \psi\left(\sqrt{\frac{r_{\boldsymbol{\gamma}_2(W_i)}(Y_i)}{r_{\boldsymbol{\gamma}_1(W_i)}(Y_i)}}\right)\right]\right|\right] \\
=&\mathbb{E}\left[\sup_{f \in \mathscr{F}_y(m_1,m_2)} \left|\sum_{i=1}^{n} \left(f(X_i) - \mathbb{E}\left[f(X_i)\right]\right)\right|\right] \\
\leq& 9.77y\sqrt{\overline{V} L_n(y)} + 90\overline{V} L_n(y),
\end{aligned} \tag{4.6.4}$$

where $L_n(y) = 9.11 + \log_+\left[n/\left(3\sqrt{2}y^2\right)\right]$. Under Assumption 4.2.2, there exists a constant $\alpha \geq 1$ such that

$$\overline{V} = 2|m_1 \vee m_2| + 1 \leq 2\alpha(|m_1| + |m_2|) + 1 \leq \left(2\alpha + \frac{1}{2}\right)(|m_1| + |m_2|). \tag{4.6.5}$$

Therefore, combining (4.6.4) and (4.6.5), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\sup_{(\mathbf{P}_{\boldsymbol{\gamma}_1},\mathbf{P}_{\boldsymbol{\gamma}_2}) \in \mathscr{B}^{\mathscr{P}_{m_1} \times \mathscr{P}_{m_2}}(\mathbf{P}^\star, y)} |\mathbf{Z}(\mathbf{X}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)|\right] \\
\leq& 9.77y\sqrt{\left(2\alpha + \frac{1}{2}\right)(|m_1| + |m_2|)\, L_n(y)} + 90\left(2\alpha + \frac{1}{2}\right)(|m_1| + |m_2|) L_n(y).
\end{aligned} \tag{4.6.6}$$

Recall that $D_n(m) = |m|\left[9.11 + \log_+(n/|m|)\right]$. For any $\eta \geq 1$, provided the condition $y^2 \geq \eta\left[D_n(m_1) + D_n(m_2)\right]$, on the one hand, we have

$$\begin{aligned} y^2 &\geq \eta\left[|m_1|\left(9.11 + \log_+\left(\frac{n}{|m_1|+|m_2|}\right)\right) + |m_2|\left(9.11 + \log_+\left(\frac{n}{|m_1|+|m_2|}\right)\right)\right] \\ &= \eta(|m_1|+|m_2|)\left[9.11 + \log_+\left(\frac{n}{|m_1|+|m_2|}\right)\right]. \end{aligned} \tag{4.6.7}$$

On the other hand, (4.6.7) also implies $y^2 \geq |m_1| + |m_2|$. Therefore,

$$\begin{aligned} L_n(y) &= 9.11 + \log_+\left(\frac{n}{3\sqrt{2}y^2}\right) \leq 9.11 + \log_+\left[\frac{n}{3\sqrt{2}(|m_1|+|m_2|)}\right] \\ &\leq 9.11 + \log_+\left(\frac{n}{|m_1|+|m_2|}\right). \end{aligned} \tag{4.6.8}$$

Plugging (4.6.7) and (4.6.8) into (4.6.6), we complete the proof. □

4.6.2 Proof of Theorem 4.2.1

The proof of Theorem 4.2.1 is inspired by the proof of Theorem A.1 in Baraud and Birgé (2018). Before we start to prove Theorem 4.2.1, we first show the following result.

Proposition 4.6.3. *Let numbers $a, \eta \geq 1$ and $\delta, \vartheta > 1$ such that*

$$2\exp(-\vartheta) + \sum_{j=1}^{+\infty}\exp(-\vartheta\delta^j) \leq 1. \tag{4.6.9}$$

Under Assumption 4.2.1 and 4.2.2, for any $\xi > 0$ and for all $m_1, m_2 \in \mathcal{M}$ simultaneously, with probability at least $1 - \Sigma^2 e^{-\xi}$,

$$\begin{aligned} &\sup_{(\mathbf{P}_{\boldsymbol{\gamma}_1}, \mathbf{P}_{\boldsymbol{\gamma}_2}) \in \mathscr{P}_{m_1} \times \mathscr{P}_{m_2}} \left[|\mathbf{Z}(\mathbf{X}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)| - k_1\left[\mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\boldsymbol{\gamma}_1}) + \mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\boldsymbol{\gamma}_2})\right]\right] \\ &\leq k_0 a\left\{\eta\left[D_n(m_1) + D_n(m_2)\right] \vee \left(\Delta(m_1) + \Delta(m_2) + \vartheta + \xi\right)\right\}, \end{aligned}$$

where

$$\begin{aligned} k_0 =& 16\sqrt{\frac{9.77\sqrt{\frac{2\alpha+1/2}{\eta}} + \frac{90(2\alpha+1/2)}{\eta} + \frac{3\sqrt{2}}{16}}{2a}} + \frac{4}{a} \\ &+ \left(9.77\sqrt{\frac{2\alpha+1/2}{\eta}} + \frac{90(2\alpha+1/2)}{\eta}\right), \end{aligned}$$

$$\begin{aligned} k_1 =& 16\sqrt{\frac{\delta\left(9.77\sqrt{\frac{2\alpha+1/2}{\eta}} + \frac{90(2\alpha+1/2)}{\eta} + \frac{3\sqrt{2}}{16}\right)}{2a}} + \frac{4}{a} \\ &+ \left(9.77\sqrt{\frac{2\alpha+1/2}{\eta}} + \frac{90(2\alpha+1/2)}{\eta}\right)\delta. \end{aligned}$$

Proof. Let $\xi > 0$, $\delta, \vartheta > 1$, $a, \eta \geq 1$ and $m_1, m_2 \in \mathcal{M}$ be fixed. For each $j \in \mathbb{N}$, we set

$$x_0(m_1, m_2) = \eta \left(D_n(m_1) + D_n(m_2) \right) \vee \left(\Delta(m_1) + \Delta(m_2) + \vartheta + \xi \right),$$

$$x_j(m_1, m_2) = \delta^j x_0(m_1, m_2), \qquad y_j^2(m_1, m_2) = a x_j(m_1, m_2).$$

For each $j \in \mathbb{N}$, we define the set

$$\begin{aligned} &\mathcal{B}_j^{\mathscr{P}_{m_1} \times \mathscr{P}_{m_2}}(\mathbf{P}^\star) \\ &= \left\{ (\mathbf{P}_{\boldsymbol{\gamma}_1}, \mathbf{P}_{\boldsymbol{\gamma}_2}) \in \mathscr{P}_{m_1} \times \mathscr{P}_{m_2} \middle| \; y_j^2 \leq \mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\boldsymbol{\gamma}_1}) + \mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\boldsymbol{\gamma}_2}) < y_{j+1}^2 \right\} \end{aligned}$$

and set

$$\mathcal{Z}_j^{\mathscr{P}_{m_1} \times \mathscr{P}_{m_2}}(\mathbf{X}) = \sup_{(\mathbf{P}_{\boldsymbol{\gamma}_1}, \mathbf{P}_{\boldsymbol{\gamma}_2}) \in \mathcal{B}_j^{\mathscr{P}_{m_1} \times \mathscr{P}_{m_2}}(\mathbf{P}^\star)} \left| \mathbf{Z}(\mathbf{X}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) \right|.$$

For simplifying the notations, let us drop the dependancy of x_j and y_j with respect to m_1, m_2 for a while. Since $\mathcal{B}_j^{\mathscr{P}_{m_1} \times \mathscr{P}_{m_2}}(\mathbf{P}^\star) \subset \mathscr{B}^{\mathscr{P}_{m_1} \times \mathscr{P}_{m_2}}(\mathbf{P}^\star, y_{j+1})$ and $y_{j+1}^2 > y_0^2 = a x_0 \geq \eta \left[D_n(m_1) + D_n(m_2) \right]$, under Assumption 4.2.2, applying Proposition 4.6.2 yields,

$$\begin{aligned} \mathbb{E}\left[\mathcal{Z}_j^{\mathscr{P}_{m_1} \times \mathscr{P}_{m_2}}(\mathbf{X}) \right] &= \mathbb{E}\left[\sup_{(\mathbf{P}_{\boldsymbol{\gamma}_1}, \mathbf{P}_{\boldsymbol{\gamma}_2}) \in \mathcal{B}_j^{\mathscr{P}_{m_1} \times \mathscr{P}_{m_2}}(\mathbf{P}^\star)} \left| \mathbf{Z}(\mathbf{X}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) \right| \right] \\ &\leq \mathbb{E}\left[\sup_{(\mathbf{P}_{\boldsymbol{\gamma}_1}, \mathbf{P}_{\boldsymbol{\gamma}_2}) \in \mathscr{B}^{\mathscr{P}_{m_1} \times \mathscr{P}_{m_2}}(\mathbf{P}^\star, y_{j+1})} \left| \mathbf{Z}(\mathbf{X}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) \right| \right] \\ &\leq \left(9.77 \sqrt{\frac{2\alpha + 1/2}{\eta}} + \frac{90(2\alpha + 1/2)}{\eta} \right) y_{j+1}^2. \end{aligned} \tag{4.6.10}$$

For $i \in \{1, \ldots, n\}$, we set

$$U_{i,(r_{\boldsymbol{\gamma}_1}, r_{\boldsymbol{\gamma}_2})} = \psi\left(\sqrt{\frac{r_{\boldsymbol{\gamma}_2(W_i)}(Y_i)}{r_{\boldsymbol{\gamma}_1(W_i)}(Y_i)}} \right) - \mathbb{E}\left[\psi\left(\sqrt{\frac{r_{\boldsymbol{\gamma}_2(W_i)}(Y_i)}{r_{\boldsymbol{\gamma}_1(W_i)}(Y_i)}} \right) \right]. \tag{4.6.11}$$

With the fact that ψ takes values in $[-1, 1]$, it is easy to observe that

$$\max_{i=1,\ldots,n} \left| U_{i,(r_{\boldsymbol{\gamma}_1}, r_{\boldsymbol{\gamma}_2})} \right| \leq 2.$$

Moreover, ψ satisfies the Assumption 2 more precisely (11) in Baraud and Birgé (2018) with $a_2^2 = 3\sqrt{2}$, we derive for each $j \in \mathbb{N}$, all $\boldsymbol{\gamma}_1 \in \boldsymbol{\Gamma}_{m_1}$, $\boldsymbol{\gamma}_2 \in \boldsymbol{\Gamma}_{m_2}$ such that $(\mathbf{P}_{\boldsymbol{\gamma}_1}, \mathbf{P}_{\boldsymbol{\gamma}_2}) \in \mathcal{B}_j^{\mathscr{P}_{m_1} \times \mathscr{P}_{m_2}}(\mathbf{P}^\star)$

$$\sum_{i=1}^n \mathbb{E}\left[U_{i,(r_{\boldsymbol{\gamma}_1}, r_{\boldsymbol{\gamma}_2})}^2 \right] \leq \sum_{i=1}^n \mathbb{E}\left[\psi^2\left(\sqrt{\frac{r_{\boldsymbol{\gamma}_2(W_i)}(Y_i)}{r_{\boldsymbol{\gamma}_1(W_i)}(Y_i)}} \right) \right] \leq 3\sqrt{2} y_{j+1}^2.$$

Then, for each $j \in \mathbb{N}$, we can apply Proposition 4.6.1 with $b = 2$, $v^2 = 3\sqrt{2}y_{j+1}^2$ and $T = \mathcal{B}_j^{\mathscr{P}_{m_1}\times\mathscr{P}_{m_2}}(\mathbf{P}^\star)$ and obtain that for all $c > 0$ and for all $(\mathbf{P}_{\boldsymbol{\gamma}_1}, \mathbf{P}_{\boldsymbol{\gamma}_2}) \in \mathcal{B}_j^{\mathscr{P}_{m_1}\times\mathscr{P}_{m_2}}(\mathbf{P}^\star)$ with probability at least $1 - e^{-x_j}$,

$$
\begin{aligned}
&|\mathbf{Z}(\mathbf{X}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)| \\
\leq &\mathcal{Z}_j^{\mathscr{P}_{m_1}\times\mathscr{P}_{m_2}}(\mathbf{X}) \\
\leq &(1+c)\mathbb{E}\left[\mathcal{Z}_j^{\mathscr{P}_{m_1}\times\mathscr{P}_{m_2}}(\mathbf{X})\right] + \frac{3\sqrt{2}y_{j+1}^2 c}{16} + 4\left(1+\frac{8}{c}\right)x_j \\
\leq &(1+c)\left(9.77\sqrt{\frac{2\alpha+1/2}{\eta}} + \frac{90(2\alpha+1/2)}{\eta}\right)y_{j+1}^2 + \frac{3\sqrt{2}y_{j+1}^2 c}{16} + 4\left(1+\frac{8}{c}\right)x_j \\
\leq &(1+c)\left(9.77\sqrt{\frac{2\alpha+1/2}{\eta}} + \frac{90(2\alpha+1/2)}{\eta}\right)y_{j+1}^2 + \frac{3\sqrt{2}y_{j+1}^2 c}{16} + \frac{4}{a}\left(1+\frac{8}{c}\right)y_j^2 \\
\leq &\left[(1+c)\left(9.77\sqrt{\frac{2\alpha+1/2}{\eta}} + \frac{90(2\alpha+1/2)}{\eta}\right)\delta + \frac{3\sqrt{2}c\delta}{16} + \frac{4}{a}\left(1+\frac{8}{c}\right)\right]y_j^2.
\end{aligned}
$$

Taking

$$
c = \sqrt{\frac{32}{\left(9.77\sqrt{\frac{2\alpha+1/2}{\eta}} + \frac{90(2\alpha+1/2)}{\eta} + \frac{3\sqrt{2}}{16}\right)\delta a}}
$$

to minimize the bracketed term yields for all $(\mathbf{P}_{\boldsymbol{\gamma}_1}, \mathbf{P}_{\boldsymbol{\gamma}_2}) \in \mathcal{B}_j^{\mathscr{P}_{m_1}\times\mathscr{P}_{m_2}}(\mathbf{P}^\star)$, with probability at least $1 - e^{-x_j}$

$$
|\mathbf{Z}(\mathbf{X}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)| \leq k_1 y_j^2.
$$

By the definition of $\mathcal{B}_j^{\mathscr{P}_{m_1}\times\mathscr{P}_{m_2}}(\mathbf{P}^\star)$, we get for all $(\mathbf{P}_{\boldsymbol{\gamma}_1}, \mathbf{P}_{\boldsymbol{\gamma}_2}) \in \mathcal{B}_j^{\mathscr{P}_{m_1}\times\mathscr{P}_{m_2}}(\mathbf{P}^\star)$, with probability at least $1 - e^{-x_j}$,

$$
|\mathbf{Z}(\mathbf{X}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)| \leq k_1 y_j^2 \leq k_1\left[\mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\boldsymbol{\gamma}_1}) + \mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\boldsymbol{\gamma}_2})\right].
$$

We define

$$
\mathcal{Z}^{\mathscr{P}_{m_1}\times\mathscr{P}_{m_2}}(\mathbf{X}) = \sup_{(\mathbf{P}_{\boldsymbol{\gamma}_1}, \mathbf{P}_{\boldsymbol{\gamma}_2}) \in \mathscr{B}^{\mathscr{P}_{m_1}\times\mathscr{P}_{m_2}}(\mathbf{P}^\star, y_0)} |\mathbf{Z}(\mathbf{X}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)|.
$$

With an analogous argument by applying Proposition 4.6.1 to $\mathcal{Z}^{\mathscr{P}_{m_1}\times\mathscr{P}_{m_2}}(\mathbf{X})$ with $x = x_0$, we can obtain for all $(\mathbf{P}_{\boldsymbol{\gamma}_1}, \mathbf{P}_{\boldsymbol{\gamma}_2}) \in \mathscr{B}^{\mathscr{P}_{m_1}\times\mathscr{P}_{m_2}}(\mathbf{P}^\star, y_0)$ and all $c > 0$, with probability at least $1 - e^{-x_0}$,

$$
\begin{aligned}
&|\mathbf{Z}(\mathbf{X}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)| \\
\leq &\mathcal{Z}^{\mathscr{P}_{m_1}\times\mathscr{P}_{m_2}}(\mathbf{X}) \\
\leq &\left[(1+c)\left(9.77\sqrt{\frac{2\alpha+1/2}{\eta}} + \frac{90(2\alpha+1/2)}{\eta}\right) + \frac{3\sqrt{2}c}{16} + \frac{4}{a}\left(1+\frac{8}{c}\right)\right]y_0^2.
\end{aligned}
$$

To minimize the bracketed term, we take

$$c = \sqrt{\frac{32}{\left(9.77\sqrt{\frac{2\alpha+1/2}{\eta}} + \frac{90(2\alpha+1/2)}{\eta} + \frac{3\sqrt{2}}{16}\right)a}}$$

and therefore for all $(\mathbf{P}_{\boldsymbol{\gamma}_1}, \mathbf{P}_{\boldsymbol{\gamma}_2}) \in \mathscr{B}^{\mathscr{P}_{m_1}\times\mathscr{P}_{m_2}}(\mathbf{P}^\star, y_0)$ with probability at least $1 - e^{-x_0}$,

$$|\mathbf{Z}(\mathbf{X}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)| \le \mathcal{Z}^{\mathscr{P}_{m_1}\times\mathscr{P}_{m_2}}(\mathbf{X}) \le a k_0 x_0.$$

Combining all the bounds together, we derive for all $(\mathbf{P}_{\boldsymbol{\gamma}_1}, \mathbf{P}_{\boldsymbol{\gamma}_2}) \in \mathscr{P}_{m_1} \times \mathscr{P}_{m_2}$ simultaneously with probability at least $1 - \varepsilon(m_1, m_2)$,

$$|\mathbf{Z}(\mathbf{X}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)| \le k_1 \left[\mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\boldsymbol{\gamma}_1}) + \mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\boldsymbol{\gamma}_2})\right] + a k_0 x_0(m_1, m_2),$$

where

$$\varepsilon(m_1, m_2) = 2\exp\left[-x_0(m_1, m_2)\right] + \sum_{j\ge 1} \exp\left[-x_j(m_1, m_2)\right].$$

By the definition of $x_j(m_1, m_2)$, we notice that for all $j \in \mathbb{N}$, $x_j(m_1, m_2) \ge \Delta(m_1) + \Delta(m_2) + \vartheta\delta^j + \xi$. Hence, provided (4.6.9), we have

$$\begin{aligned}\varepsilon(m_1, m_2) &\le \exp\left[-\xi - \Delta(m_1) - \Delta(m_2)\right]\left(2\exp(-\vartheta) + \sum_{j\ge 1}\exp(-\vartheta\delta^j)\right)\\ &\le \exp\left[-\xi - \Delta(m_1) - \Delta(m_2)\right].\end{aligned}$$

Finally we can extend this result to all $(\mathbf{P}_{\boldsymbol{\gamma}_1}, \mathbf{P}_{\boldsymbol{\gamma}_2}) \in \mathscr{P} \times \mathscr{P}$ by summing these bounds over $(m_1, m_2) \in \mathcal{M} \times \mathcal{M}$ and using (4.2.1). □

Proof of Theorem 4.2.1. We apply Proposition 4.6.3 with $\delta = 1.175$, $\vartheta = 1.47$ and as for the values of η and a, we shall choose them later such that $k_1 = 3\beta/8$, with some $0 < \beta < 1$. On a set Ω_ξ the probability of which is at least $1 - \Sigma^2 e^{-\xi}$, for all $\mathbf{P}_{\boldsymbol{\gamma}_1}, \mathbf{P}_{\boldsymbol{\gamma}_2} \in \mathscr{P}$ and all $\mathscr{P}_{m_1} \times \mathscr{P}_{m_2}$ containing $(\mathbf{P}_{\boldsymbol{\gamma}_1}, \mathbf{P}_{\boldsymbol{\gamma}_2})$

$$\begin{aligned}\mathbf{T}(\mathbf{X}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) &\le \mathbb{E}\left[\mathbf{T}(\mathbf{X}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)\right] + \frac{3\beta}{8}\left[\mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\boldsymbol{\gamma}_1}) + \mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\boldsymbol{\gamma}_2})\right]\\ &\quad + k_0 a\left[\eta\left(D_n(m_1) + D_n(m_2)\right) \vee \left(\Delta(m_1) + \Delta(m_2) + \vartheta + \xi\right)\right]\\ &\le \mathbb{E}\left[\mathbf{T}(\mathbf{X}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)\right] + \frac{3\beta}{8}\left[\mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\boldsymbol{\gamma}_1}) + \mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\boldsymbol{\gamma}_2})\right]\\ &\quad + k_0 a\left[\eta D_n(m_1) + \eta D_n(m_2) + \Delta(m_1) + \Delta(m_2) + \vartheta + \xi\right].\end{aligned}$$

Since the last inequality is true for all the $\mathscr{P}_{m_1} \times \mathscr{P}_{m_2}$ containing $(\mathbf{P}_{\boldsymbol{\gamma}_1}, \mathbf{P}_{\boldsymbol{\gamma}_2})$, provided $C_0(2\alpha + 1/2) \ge k_0 a\eta$, we derive from (4.2.2) that with a probability at least $1 - \Sigma^2 e^{-\xi}$,

$$\begin{aligned}\mathbf{T}(\mathbf{X}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) &\le \mathbb{E}\left[\mathbf{T}(\mathbf{X}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)\right] + \frac{3\beta}{8}\left[\mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\boldsymbol{\gamma}_1}) + \mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\boldsymbol{\gamma}_2})\right]\\ &\quad + \mathbf{pen}(\boldsymbol{\gamma}_1) + \mathbf{pen}(\boldsymbol{\gamma}_2) + k_0 a(\vartheta + \xi). \qquad (4.6.12)\end{aligned}$$

According to Proposition 3 of Baraud and Birgé (2018), the function ψ satisfies Assumption 2 (more precisely (10)) in the same paper with $a_0 = 4$ and $a_1 = 3/8$. As a consequence, for all $\mathbf{P}_{\gamma_1}, \mathbf{P}_{\gamma_2} \in \mathscr{P}$ and $\mathbf{P}^\star \in \mathscr{P}^f$,

$$\mathbb{E}\left[\mathbf{T}(\mathbf{X}, \gamma_1, \gamma_2)\right] \leq 4\mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\gamma_1}) - \frac{3}{8}\mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\gamma_2}). \tag{4.6.13}$$

Combining (4.6.12) and (4.6.13), we derive that for all $\mathbf{P}_{\gamma_1}, \mathbf{P}_{\gamma_2} \in \mathscr{P}$ and $\mathbf{P}^\star \in \mathscr{P}^f$, with a probability at least $1 - \Sigma^2 e^{-\xi}$,

$$\begin{aligned}\mathbf{T}(\mathbf{X}, \gamma_1, \gamma_2) \leq (4 + \frac{3\beta}{8})\mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\gamma_1}) - \frac{3(1-\beta)}{8}\mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\gamma_2}) \\ + \mathbf{pen}(\gamma_1) + \mathbf{pen}(\gamma_2) + k_0 a(\vartheta + \xi).\end{aligned} \tag{4.6.14}$$

This entails that, for any (random) elements $\mathbf{P}_{\widehat{\gamma}_\lambda}, \mathbf{P}_{\widehat{\gamma}_{\widehat{\lambda}}} \in \mathscr{P}$, on a set Ω_ξ with probability at least $1 - \Sigma^2 e^{-\xi}$

$$\begin{aligned}\mathbf{T}(\mathbf{X}, \widehat{\gamma}_\lambda, \widehat{\gamma}_{\widehat{\lambda}}) \leq (4 + \frac{3\beta}{8})\mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\widehat{\gamma}_\lambda}) - \frac{3(1-\beta)}{8}\mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\widehat{\gamma}_{\widehat{\lambda}}}) \\ + \mathbf{pen}(\widehat{\gamma}_\lambda) + \mathbf{pen}(\widehat{\gamma}_{\widehat{\lambda}}) + k_0 a(\vartheta + \xi)\end{aligned} \tag{4.6.15}$$

and

$$\begin{aligned}\boldsymbol{v}(\boldsymbol{X}, \widehat{\gamma}_\lambda) &= \sup_{\lambda' \in \Lambda} \left[\mathbf{T}(\boldsymbol{X}, \widehat{\gamma}_\lambda, \widehat{\gamma}_{\lambda'}) - \mathbf{pen}(\widehat{\gamma}_{\lambda'})\right] + \mathbf{pen}(\widehat{\gamma}_\lambda) \\ &\leq (4 + \frac{3\beta}{8})\mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\widehat{\gamma}_\lambda}) - \frac{3(1-\beta)}{8} \inf_{\lambda' \in \Lambda} \mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\widehat{\gamma}_{\lambda'}}) \\ &\quad + 2\,\mathbf{pen}(\widehat{\gamma}_\lambda) + k_0 a(\vartheta + \xi).\end{aligned} \tag{4.6.16}$$

By the construction of ψ, $\mathbf{T}(\mathbf{X}, \widehat{\gamma}_{\widehat{\lambda}}, \widehat{\gamma}_\lambda) = -\mathbf{T}(\mathbf{X}, \widehat{\gamma}_\lambda, \widehat{\gamma}_{\widehat{\lambda}})$. Combining (4.6.15), (4.6.16) and (4.2.3) leads to for any $\lambda \in \Lambda$, on a set Ω_ξ with probability at least $1 - \Sigma^2 e^{-\xi}$

$$\begin{aligned}\frac{3(1-\beta)}{8}\mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\widehat{\gamma}_{\widehat{\lambda}}}) &\leq (4 + \frac{3\beta}{8})\mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\widehat{\gamma}_\lambda}) - \mathbf{T}(\mathbf{X}, \widehat{\gamma}_\lambda, \widehat{\gamma}_{\widehat{\lambda}}) \\ &\quad + \mathbf{pen}(\widehat{\gamma}_\lambda) + \mathbf{pen}(\widehat{\gamma}_{\widehat{\lambda}}) + k_0 a(\vartheta + \xi) \\ &\leq (4 + \frac{3\beta}{8})\mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\widehat{\gamma}_\lambda}) + \left[\mathbf{T}(\mathbf{X}, \widehat{\gamma}_{\widehat{\lambda}}, \widehat{\gamma}_\lambda) - \mathbf{pen}(\widehat{\gamma}_\lambda)\right] \\ &\quad + \mathbf{pen}(\widehat{\gamma}_{\widehat{\lambda}}) + 2\,\mathbf{pen}(\widehat{\gamma}_\lambda) + k_0 a(\vartheta + \xi) \\ &\leq (4 + \frac{3\beta}{8})\mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\widehat{\gamma}_\lambda}) + \boldsymbol{v}(\boldsymbol{X}, \widehat{\gamma}_{\widehat{\lambda}}) + 2\,\mathbf{pen}(\widehat{\gamma}_\lambda) + k_0 a(\vartheta + \xi) \\ &\leq (4 + \frac{3\beta}{8})\mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\widehat{\gamma}_\lambda}) + \boldsymbol{v}(\boldsymbol{X}, \widehat{\gamma}_\lambda) + 1 + 2\,\mathbf{pen}(\widehat{\gamma}_\lambda) + k_0 a(\vartheta + \xi).\end{aligned} \tag{4.6.17}$$

Plugging (4.6.16) into (4.6.17) yields, for any $\lambda \in \Lambda$, on a set Ω_ξ with probability at least $1 - \Sigma^2 e^{-\xi}$,

$$\frac{3(1-\beta)}{8}\mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\widehat{\gamma}_{\widehat{\lambda}}}) \leq (8 + \frac{3\beta}{4})\mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\widehat{\gamma}_\lambda}) + 4\,\mathbf{pen}(\widehat{\gamma}_\lambda) + 2k_0 a(\vartheta + \xi) + 1.$$

Therefore, for any $\lambda \in \Lambda$ on a set Ω_ξ with probability at least $1 - \Sigma^2 e^{-\xi}$,

$$\mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}}) \leq \frac{64+6\beta}{3(1-\beta)} \mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\widehat{\boldsymbol{\gamma}}_\lambda}) + \frac{32}{3(1-\beta)} \mathbf{pen}(\widehat{\boldsymbol{\gamma}}_\lambda) + \frac{16 k_0 a(\vartheta+\xi)+8}{3(1-\beta)}. \quad (4.6.18)$$

By the equality (4.6.2), we rewrite (4.6.18) as the following

$$\mathbf{h}^2(\mathbf{Q}^\star, \mathbf{R}_{\widehat{\boldsymbol{\gamma}}_{\widehat{\lambda}}}) \leq \frac{64+6\beta}{3(1-\beta)} \mathbf{h}^2(\mathbf{Q}^\star, \mathbf{R}_{\widehat{\boldsymbol{\gamma}}_\lambda}) + \frac{32}{3(1-\beta)} \mathbf{pen}(\widehat{\boldsymbol{\gamma}}_\lambda) + \frac{16 k_0 a(\vartheta+\xi)+8}{3(1-\beta)}. \quad (4.6.19)$$

Taking $\beta = 0.75$, $\eta \approx 9947.13(2\alpha + 1/2)$, we can compute the value of $a \approx 2365.57$ such that $k_1 = 3\beta/8$ and $k_0 \approx 0.251$. Therefore, provided $C_0 \geq 5.9 \times 10^6$, plugging the values of β, k_0, a and ϑ into (4.6.19), we finally conclude. □

4.7 Signals for testing Poisson and exponential models

`fms-type` (Poisson): $n = 497$, changepoints are located at the positions

$$l_0 = \left(\frac{139}{497}, \frac{226}{497}, \frac{243}{497}, \frac{300}{497}, \frac{309}{497}, \frac{333}{497}\right).$$

The Poisson mean on each segment is 4, 6, 10, 3, 7, 1, 5 respectively, i.e. $\boldsymbol{\gamma}^\star$ takes the value $\log 4$, $\log 6$, $\log 10$, $\log 3$, $\log 7$, $\log 1$, $\log 5$ on each segment. For this signal, we also test the scenario when outliers present in the observations by randomly modifying five points in the observations into 30.

`mix-type` (Poisson): $n = 560$ and $\boldsymbol{\gamma}^\star$ is a piecewise constant function on $[0, 1)$ with 13 changepoints at a sequence of locations

$$l_0 = \left(\frac{11}{560}, \frac{21}{560}, \frac{41}{560}, \frac{61}{560}, \frac{91}{560}, \frac{121}{560}, \frac{161}{560}, \frac{201}{560}, \frac{251}{560}, \frac{301}{560}, \frac{361}{560}, \frac{421}{560}, \frac{491}{560}\right)$$

and on each segment the Poisson mean $e^{\boldsymbol{\gamma}^\star}$ is given by the value 30, 2, 26, 4, 24, 6, 22, 8, 20, 10, 18, 12, 16, 14 respectively.

`teeth-type` (exponential): $n = 140$ and $\boldsymbol{\gamma}^\star$ is a piecewise constant function on $[0, 1)$ with 13 changepoints at a sequence of locations

$$l_0 = \left(\frac{11}{140}, \frac{21}{140}, \frac{31}{140}, \frac{41}{140}, \frac{51}{140}, \frac{61}{140}, \frac{71}{140}, \frac{81}{140}, \frac{91}{140}, \frac{101}{140}, \frac{111}{140}, \frac{121}{140}, \frac{131}{140}\right)$$

and on each segment the value of $\boldsymbol{\gamma}^\star$ is given by 0.5, 5, 0.5, 5, 0.5, 5, 0.5, 5, 0.5, 5, 0.5, 5, 0.5, 5 respectively. For this signal, we also test the scenario when outliers present in the observations by randomly modifying two points in the observations into 20.

`stairs-type` (exponential): $n = 500$ and $\boldsymbol{\gamma}^\star$ is a piecewise constant function on $[0, 1)$ with 4 changepoints at a sequence of locations

$$l_0 = \left(\frac{101}{500}, \frac{201}{500}, \frac{301}{500}, \frac{401}{500}\right)$$

and on each segment the value of $\boldsymbol{\gamma}^\star$ is given by 2^4, 2^2, 1, 2^{-2}, 2^{-4} respectively.

Bibliography

Akakpo, N. (2012). Adaptation to anisotropy and inhomogeneity via dyadic piecewise polynomial selection. *Math. Methods Statist.*, **21**, 1–28.

Albertson, D. G. and Pinkel, D. (2003). Genomic microarrays in human genetic disease and cancer. *Hum. Mol. Genet.*, **12**, 145–152.

Anscombe, F. J. (1948). The transformation of Poisson, binomial and negative-binomial data. *Biometrika*, **35**, 246–254.

Antoniadis, A. and Leblanc, F. (2000). Nonparametric wavelet regression for binary response. *J. Theor. Appl. Stat.*, **34**, 183–213.

Antoniadis, A. and Sapatinas, T. (2001). Wavelet shrinkage for natural exponential families with quadratic variance functions. *Biometrika*, **88**, 805–820.

Antoniadis, A., Besbeas, P. and Sapatinas, T. (2001). Wavelet shrinkage for natural exponential families with cubic variance functions. *Sankhyā: Indian J. Stat., Ser. A*, **63**, 309–327.

Baraud, Y. (2002). Model selection for regression on a random design. *ESAIM Probab. Stat.*, **6**, 127–146.

Baraud, Y. (2021). Tests and estimation strategies associated to some loss functions. *Probab. Theory Related Fields*, **180**, 799–846.

Baraud, Y. and Birgé, L. (2009). Estimating the intensity of a random measure by histogram type estimators. *Probab. Theory Related Fields*, **143**, 239–284.

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

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

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

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

Baraud, Y., Giraud, C. and Huet, S. (2009). Gaussian model selection with an unknown variance. *Ann. Statist.*, **37**, 630–672.

Baraud, Y., Giraud, C. and Huet, S. (2014). Estimator selection in the Gaussian setting. *Ann. Inst. H. Poincaré Probab. Statist.*, **50**, 1092–1119.

Barron, A., Birgé, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, **113**, 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. *J. Mach. Learn. Res.*, **20**, 1–17.

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

Beran, R. and Dümbgen, L. (1998). Modulation of estimators and confidence sets. *Ann. Statist.*, **26**, 1826–1856.

Besbeas, P., De Feis, I. and Sapatinas, T. (2004). A comparative simulation study of wavelet shrinkage estimators for Poisson counts. *Internat. Statist. Rev.*, **72**, 209–237.

Birgé, L. (1986). On estimating a density using Hellinger distance and some other strange facts. *Probab. Theory Related Fields*, **71**, 271–291.

Birgé, L. (2006). Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Ann. Inst. H. Poincaré Probab. Stat.*, **42**, 273–325.

Birgé, L. and Massart, P. (1997). From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, 55–87. Springer, New York.

Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc.*, **3**, 203–268.

Blythe, D. A. J., von Bunau, P., Meinecke, F. C. and Muller, K.-R. (2012). Feature extraction for change-point detection using stationary subspace analysis. *IEEE Trans. Neural Netw. Learn. Syst.*, **23**, 631–643.

Breiman, L., Friedman, J., Stone, C. J. and Olshen, R. A. (1984). *Classification and Regression Trees*. Taylor & Francis, New York.

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

Bunea, F., Tsybakov, A. B., and Wegkamp, M. H. (2007). Aggregation for Gaussian regression. *Ann. Statist.*, **35**, 1674–1697.

Candès, E. and Sur, P. (2020). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *Ann. Statist.*, **48**, 27–42.

Chen, J. (2022). Estimating a regression function in exponential families by model selection. *arXiv preprint*, arXiv:2203.06656.

Cleynen, A. and Lebarbier, E. (2014). Segmentation of the Poisson and negative binomial rate models: a penalized estimator. *ESAIM Probab. Stat.*, **18**, 750–769.

Cleynen, A. and Lebarbier, E. (2017). Model selection for the segmentation of multiparameter exponential family distributions. *Electron. J. Stat.*, **11**, 800–842.

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

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

DeVore, R. A. and Lorentz, G. G. (1993). *Constructive Approximation.* Springer-Verlag, New York.

Donoho, D. L. (1993). Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data. In *Proceedings of Symposia in Applied Mathematics: Different Perspectives on Wavelets* (Daubechies, I. ed.), **47**, 173–205.

Donoho, D. L. and Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.*, **26**, 879–921.

Fearnhead, P. (2006). Exact and efficient Bayesian inference for multiple changepoint. *Statist. Comput.*, **16**, 203–213.

Fearnhead, P. and Rigaill, G. (2019). Changepoint detection in the presence of outliers. *J. Amer. Statist. Assoc.*, **114**, 169–183.

Fearnhead, P. and Rigaill, G. (2020). Relating and comparing methods for detecting changes in mean. *Stat*, e291.

Fisz, M. (1955). The limiting distribution of a function of two independent random variables and its statistical application. *Colloq. Math.*, **3**, 138–146.

Frick, K., Munk, A. and Sieling, H. (2013). Multiscale change point inference. *J. Roy. Statist. Soc., Ser. B*, **76**, 495–580.

Fryźlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *Ann. Statist.*, **42**, 2243–2281.

Fryźlewicz, P. and Nason, G. P. (2001). Poisson intensity estimation using wavelets and the Fisz transformation. Technical Report, **01/10**. Department of Mathematics, University of Bristol, United Kingdom.

Fryźlewicz, P. and Nason, G. P. (2004). A Haar-Fisz algorithm for Poisson intensity estimation. *J. Comput. Graph. Statist.*, **13**, 621–638.

Gallagher, C., Lund, R. and Robbins, M. (2013). Changepoint detection in climate time series with long-term trends. *J. Clim.*, **26**, 4994–5006.

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

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Hansen, N. (2016). The CMA evolution strategy: a tutorial. *arXiv preprint*, arXiv:1604.00772.

Haussler, D. (1995). Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *J. Combin. Theory Ser. A*, **69**, 217–232.

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

Horowitz, J. L. and Mammen, E. (2007). Rate-optimal estimation for a general class of nonparametric regression models with unknown link functions. *Ann. Statist.*, **35**, 2589–2619.

Hotz, T., Schütte, O. M., Sieling, H., Polupanow, T., Diederichsen, U., Steinem, C. and Munk, A. (2013). Idealizing ion channel recordings by a jump segmentation multiresolution filter. *IEEE Trans. NanoBioscience*, **12**, 376–386.

Huang, T., Wu, B., Lizardi, P. and Zhao, H. (2005). Detection of DNA copy number alterations using penalized least squares regression. *Bioinformatics*, **21**, 3811–3817.

Ibragimov, I. A. and Has'minskiĭ, R. Z. (1981). *Statistical Estimation. Asymptotic Theory*, volume 16. Springer-Verlag, New York.

Ivanoff, S., Picard, F. and Rivoirard, V. (2016). Adaptive Lasso and group-Lasso for functional Poisson regression. *J. Mach. Learn. Res.*, **17**, 1–46.

Killick, R., Fearnhead, P. and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.*, **107**, 1590–1598.

Kohler, M. (2000). Inequalities for uniform deviations of averages from expectations with applications to nonparametric regression. *J. Statist. Plann. Inference.*, **89**, 1–23.

Kroll, M. (2019). Non-parametric Poisson regression from independent and weakly dependent observations by model selection. *J. Statist. Plann. Inference*, **199**, 249–270.

Jia, J., Xie, F. and Xu, L. (2019). Sparse Poisson regression with penalized weighted score function. *Electron. J. Stat.*, **13**, 2898–2920.

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer Series in Statistics. Springer, New York.

Le Cam, L. and Yang, G. L. (1990). *Asymptotics in Statistics : Some Basic Concepts*. Springer Series in Statistics. Springer, New York.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin. Isoperimetry and processes.

Li, H., Munk, A. and Sieling, H. (2016). FDR-control in multiscale change-point segmentation. *Electron. J. Stat.*, **10**, 918–959.

Li, Y. and Cevher, V. (2015). Consistency of ℓ_1-regularized maximum-likelihood for compressive Poisson regression. In *2015 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 3606–3610.

Lloyd, C., Gunter, T., Osborne, M. A. and Roberts, S. J. (2015). Variational inference for Gaussian process modulated Poisson processes. In *International Conference on Machine Learning*, **37**, 1814–1822.

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.

Muggeo, V. M. R. and Adelfio, G. (2010). Efficient change point detection for genomic sequences of continuous measurements. *Bioinformatics*, **27**, 161–166.

Nunes, M. A. and Nason, G. P. (2009). A multiscale variance stabilization for binomial sequence proportion estimation. *Statist. Sinica*, **19**, 1491–1510.

Olshen, A. B., Venkatraman, E., Lucito, R. and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.

Pierre-Jean, M., Rigaill, G. and Neuvial, P. (2015). Performance evaluation of DNA copy number segmentation methods. *Brief. Bioinformatics*, **16**, 600–615.

Redon, R., Ishikawa, S., Fitch, K., Feuk, L., Perry, G., Andrews, T., Fiegler, H., Shapero, M., Carson, A., Chen, W., Cho, E., Dallaire, S., Freeman, J., Gonzalez, J., Gratacòs, M., Huang, J., Kalaitzopoulos, D., Komura, D., Macdonald, J. and Hurles, M. (2006). Global variation in copy number in the human genome. *Nature*, **444**, 444-454.

Reeves, J., Chen, J., Wang, X. L., Lund, R. and Lu, Q. Q. (2007). A review and comparison of changepoint detection techniques for climate data. *J. Appl. Meteorol. and Climatol.*, **46**, 900–915.

Rigaill, G. (2015). A pruned dynamic programming algorithm to recover the best segmentations with 1 to K_{max} change-points. *arXiv preprint*, arXiv:1004.0887.

Sardy, S., Antoniadis, A. and Tseng, P. (2004). Automatic smoothing with wavelets for a wide class of distributions. *J. Comput. Graph. Statist.*, **13**, 399–421.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *Ann. Statist.*, **48**, 1875–1897.

Schumaker, L. (1981). *Spline Functions: Basic Theory.* Wiley, New York.

Scott, A. J. and Knott, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, **30**, 507–512.

Spokoiny, V. (2009). Multiscale local change point detection with applications to value-at-risk. *Ann. Statist.*, **37**, 1405–1436.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, **10**, 1040–1053.

Stone, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.*, **13**, 689–705.

Suzuki, T. and Nitanda, A. (2019). Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic Besov space. *arXiv preprint*, arXiv:1910.12799.

Tibshirani, R. and Wang, P. (2007). Spatial smoothing and hot spot detection for CGH data using the fused Lasso. *Biostatistics*, **9**, 18–29.

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

Truong, C., Oudre, L. and Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Process.*, **167**, 107299.

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.

van der Vaart, A. W. and Wellner, J. A. (2009). A note on bounds for VC dimensions. In *High Dimensional Probability V: the Luminy volume*, volume 5 of *Inst. Math. Stat. Collect.*, 103–107. Inst. Math. Statist., Beachwood, OH.

Venkatraman, E. S. and Olshen, A. B. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, **23**, 657–663.

Verzelen, N., Fromont, M., Lerasle, M. and Reynaud-Bouret, P. (2020). Optimal change-point detection and localization. *arXiv preprint*, arXiv:2010.11470.

Wang, D., Yu, Y. and Rinaldo, A. (2020). Univariate mean change point detection: penalization, CUSUM and optimality. *Electron. J. Stat.*, **14**, 1917–1961.

Wegkamp, M. (2003). Model selection in nonparametric regression. *Ann. Statist.*, **31**, 252–273.

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

Yang, Y. (2004). Aggregating regression procedures to improve performance. *Bernoulli*, **10**, 25–47.

Yang, T. Y. and Kuo, L. (2001). Bayesian binary segmentation procedure for a Poisson process with multiple changepoints. *J. Comput. Graph. Statist.*, **10**, 772–785.

Zhang, N. R. and Siegmund, D. O. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, **63**, 22–32.