ROBUST ESTIMATION OF A REGRESSION FUNCTION IN EXPONENTIAL FAMILIES

YANNICK BARAUD AND JUNTONG CHEN

ABSTRACT. We observe n pairs $X_1 = (W_1, Y_1), \ldots, X_n = (W_n, Y_n)$ of independent random variables and assume, although this might not be true, that for each $i \in \{1, \ldots, n\}$, the conditional distribution of Y_i given W_i belongs to a given exponential family with real parameter $\theta_i^\star = \boldsymbol{\theta}^\star(W_i)$ the value of which is a function $\boldsymbol{\theta}^\star$ of the covariate W_i. Given a model $\overline{\boldsymbol{\Theta}}$ for $\boldsymbol{\theta}^\star$, we propose an estimator $\widehat{\boldsymbol{\theta}}$ with values in $\overline{\boldsymbol{\Theta}}$ the construction of which is independent of the distribution of the W_i and that possesses the properties of being robust to contamination, outliers and model misspecification. We establish non-asymptotic exponential inequalities for the upper deviations of a Hellinger-type distance between the true distribution of the data and the estimated one based on $\widehat{\boldsymbol{\theta}}$. Under a suitable parametrization of the exponential family, we deduce a uniform risk bound for $\widehat{\boldsymbol{\theta}}$ over the class of Hölderian functions and we prove the optimality of this bound up to a logarithmic factor. Finally, we provide an algorithm for calculating $\widehat{\boldsymbol{\theta}}$ when $\boldsymbol{\theta}^\star$ is assumed to belong to functional classes of low or medium dimensions (in a suitable sense) and, on a simulation study, we 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.

1. INTRODUCTION

In order to motivate the statistical problem we wish to solve here, let us start with a preliminary example.

Example 1. We study a cohort of n patients with respective clinical characteristics $W_1, \ldots, W_n$ 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 he/she

Date: October 28th, 2020.

2010 *Mathematics Subject Classification.* Primary 62J12, 62F35, 62G35, 62G05; Secondary 60G99.

Key words and phrases. Generalized linear model, logit (logistic) regression, Poisson regression, robust estimation, supremum of an empirical process.

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

develops the disease D and $Y_i = -1$ otherwise. A classical model for studying the effect of the clinical characteristic W on the probability of developing the disease D is the logit one

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

where $w^\star$ is an unknown vector and $\langle \cdot, \cdot \rangle$ the inner product of $\mathbb{R}^d$. The problem is to estimate $w^\star$ on the basis of the observations (W_i, Y_i) for $i \in \{1, \ldots, n\}$.

A common way of solving this problem is to use the Maximum Likelihood Estimator (MLE for short). 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 $\lambda > 0$ writes as

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

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, getting back to our Example 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). A natural question arises: how can we provide a suitable estimation of $w^\star$ despite the presence of such possible corrupted data?

This is the kind of issues 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. More precisely, we assume that we observe n pairs of independent random variables $(W_1, Y_1), \ldots, (W_n, Y_n)$ with values in $\mathscr{W} \times \mathscr{Y}$ and assume that the conditional distribution of Y_i given W_i belongs to an one-parameter exponential family with parameter $\theta_i = \boldsymbol{\theta}^\star(W_i) \in \mathbb{R}$ which is an unknown function $\boldsymbol{\theta}^\star$ of the covariate W_i for all $i \in \{1, \ldots, n\}$. Our aim is to estimate the function $\boldsymbol{\theta}^\star$ from the observations of $(W_1, Y_1), \ldots, (W_n, Y_n)$.

To our knowledge, there exist only few papers in the literature that tackle this estimation problem and establish risk bounds for the proposed estimators of $\boldsymbol{\theta}^\star$. When $\mathscr{W} = [0,1]$, Kolaczyk and Nowak (2005) proposed an estimation of $\boldsymbol{\theta}^\star$ by piecewise polynomials. When the exponential family is given in its canonical form and the natural parameter is a smooth function of the mean, they propose estimators that achieve, up to extra logarithmic factors, the classical rate $n^{-\alpha/(2\alpha+1)}$ over Besov balls with regularity $\alpha > 0$ for an Hellinger-type loss. Brown *et al* (2010) considered one-parameter exponential families which possess the property that the variances of the distributions are quadratic functions of their means. These families include as special cases the binomial, gamma and Poisson distributions, among others, and have been studied earlier by Antoniadis and Sapatinas (2001). When the exponential family is parametrized by its mean, Brown *et al* (2010) used a variance stabilizing transformation in order to turn the original problem of estimating the function $\boldsymbol{\theta}^\star$ into that of estimating a regression function in the homoscedastic Gaussian regression framework. They established uniform rates of convergence with respect to some $\mathbb{L}_2$-loss over classes of functions $\boldsymbol{\theta}^\star$ that belong to Besov balls and are bounded from above and below by positive numbers. Finally, in the case of the Poisson family parametrized by its mean, Kroll (2019) proposed an estimator of $\boldsymbol{\theta}^\star$ which is based on a model selection procedure. He proved that, under suitable but rather restrictive conditions, his estimator achieved the minimax rate of convergence over Sobolev-type ellipsoids.

We shall not tackle here the problems of model selection or adaptation. These issues will be considered in a subsequent paper. However, we shall rather focus here on the rates that can be obtained when we make a smoothness assumption on the parameter, namely that it belongs to an Hölderian class of functions with regularity $\alpha \in (0,1]$ and the risk is measured using some Hellinger-type loss. We shall first show that, if we make this smoothness assumption on a more or less arbitrary, even quite natural like the mean, parametrization of the exponential family, the optimal rate can be much slower than the classical one, i.e. $n^{-\alpha/(2\alpha+1)}$. Such a phenomenon also appears in density estimation for the Hellinger loss as proved in Birgé (1986). In order to avoid it, given an arbitrary exponential family, we shall first introduce a particular parametrization which stabilizes its Fisher information and use the Hölderian assumption for this specific parametrization. Then we shall use an estimator which is proven to remain stable, up to some extent, to the presence of outliers, contaminating data and a possible misspecification of the model. This means that its risk is the sum of two terms, one corresponding to the performance of the estimator when the parameter does belong to the model and one measuring the approximation error between the model and the true parameter. This property allows us to build the estimator on an approximate but simple model for the given Hölderian class, namely a finite dimensional linear space. With such a simple model, it is

easy to derive upper bounds for the risk using a method called ρ-estimation which has been introduced in the papers Baraud *et al* (2017) and Baraud and Birgé (2018) and which solves the problem of estimating $\boldsymbol{\theta}^\star$ under very mild assumptions.

Our main result takes the form of a non-asymptotic exponential deviation inequality for an Hellinger-type loss between the true conditional distribution of the data and the estimated one based on our model. The values of the numerical constants that are involved in this inequality are given explicitly and they do not depend on the exponential family. We shall also present an algorithm as well as a simulation study for calculating our estimator and evaluating its practical performance. Finally, let us mention that the proof of this main result relies on an upper bound for the expectation of the supremum of an empirical process over a VC-subgraph class. This bound provides explicit numerical constants, which is to our knowledge new in the literature and can be of independent interest.

The paper is organized as follows. We describe our statistical framework in Section 2 and present there several examples to which our approach applies. The construction of the estimator and our main result about its risk are presented in Section 3. We shall also explain why the deviation inequality we derive guarantees the desired robustness property of the estimator. Uniform risk bounds over Hölderian classes for a suitably chosen parameter are established in Section 4. As already mentioned, we shall see that, even when the exponential family is parametrized by the mean of the distribution, the rates we get may differ from the (classical) ones established in the Gaussian case. Section 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 6 as well as its proof. Section 7 is devoted to the other proofs.

2. The statistical setting

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 $(\mathscr{X}, \mathcal{X}) = (\mathscr{W} \times \mathscr{Y}, \mathcal{W} \otimes \mathcal{Y})$. We assume that for each $i \in \{1, \ldots, n\}$, the conditional probability of Y_i given W_i exists and is given by the value at W_i of a measurable function $Q_i^\star$ from $(\mathscr{W}, \mathcal{W})$ into the set all probabilities on $(\mathscr{Y}, \mathcal{Y})$ equipped with the Borel σ-algebra $\mathcal{T}$ associated to the total variation distance (which induces the same topology as the Hellinger one). In particular the mapping $w \mapsto h^2(Q_i^\star(w), Q)$ is measurable whatever the probability Q in $(\mathscr{Y}, \mathcal{Y})$ and $i \in \{1, \ldots, n\}$.

With a slight abuse of language, we shall also refer to $Q_i^\star$ as the conditional distribution of Y_i given W_i although this distribution is actually the value of $Q_i^\star$ at W_i. Apart from independence of the W_i, $1 \leqslant i \leqslant n$, we shall assume

nothing about their respective distributions P_{W_i} which can therefore be arbitrary.

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

$$q_\theta : y \mapsto e^{S(y)\theta - A(\theta)} \quad \text{with} \quad A(\theta) = \log\left[\int_{\mathscr{Y}} e^{\theta S(y)} d\nu(y)\right], \tag{2}$$

where S is real-valued measurable functions 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 the interior $\mathring{I}$ of 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 $\gamma = u(\theta)$ where u is a continuous and strictly monotone function on I. We shall see in Section 3.4 that our main result remains unchanged under such a transformation and we therefore choose, for the sake of simplicity, to present our statistical setting under a natural parametrization.

Given a class of functions $\overline{\boldsymbol{\Theta}}$ from $\mathscr{W}$ into I, we presume all the conditional distributions $Q_i^\star(W_i)$ are of the form $Q_{\boldsymbol{\theta}^\star(W_i)}$ with $\boldsymbol{\theta}^\star$ in $\overline{\boldsymbol{\Theta}}$. We shall refer to $\boldsymbol{\theta}^\star$ as the *regression function.*

Throughout this paper, we shall keep in mind that all these assumptions about the $Q_i^\star$ might not be true: the conditional distributions $Q_i^\star(W_i)$ might not be exactly of the form $Q_{\boldsymbol{\theta}^\star(W_i)}$ but only close to distributions of these forms and, even if they were, the set $\overline{\boldsymbol{\Theta}}$ might not contain $\boldsymbol{\theta}^\star$ but only provide a suitable approximation of it. Nevertheless, we base our construction on the assumption that the conditional probabilities $Q_i^\star$ are all equal to some element the set $\{Q_{\boldsymbol{\theta}} : w \mapsto Q_{\boldsymbol{\theta}(w)},\ \boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}}\}$ which is associated to the exponential family $\overline{\mathscr{Q}}$ and the function space $\overline{\boldsymbol{\Theta}}$.

For $i \in \{1, \ldots, n\}$, let $\mathscr{Q}_{\mathscr{W}}$ be the set of all measurable mappings from $(\mathscr{W}, \mathcal{W})$ into to the space of probabilities on $(\mathscr{Y}, \mathcal{Y})$ equipped with the topology $\mathcal{T}$, and $\boldsymbol{\mathscr{Q}}_{\mathscr{W}} = \mathscr{Q}_{\mathscr{W}}^n$. Hence, the vector $\mathbf{Q}^\star = (Q_1^\star, \ldots, Q_n^\star)$ of the true conditional distributions belongs to $\boldsymbol{\mathscr{Q}}_{\mathscr{W}}$. We endow the space $\boldsymbol{\mathscr{Q}}_{\mathscr{W}}$ with the Hellinger-type (pseudo) distance $\mathbf{h}$ defined as follows. For $\mathbf{Q} = (Q_1, \ldots, Q_n)$ and $\mathbf{Q}' = (Q_1', \ldots, Q_n')$ in $\boldsymbol{\mathscr{Q}}_{\mathscr{W}}$,

$$\begin{aligned} \mathbf{h}^2(\mathbf{Q}, \mathbf{Q}') &= \mathbb{E}\left[\sum_{i=1}^n h^2\left(Q_i(W_i), Q_i'(W_i)\right)\right] \\ &= \sum_{i=1}^n \int_{\mathscr{W}} h^2\left(Q_i(w), Q_i'(w)\right) dP_{W_i}(w) \end{aligned} \tag{3}$$

where h denotes the Hellinger distance. 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. We recall that the Hellinger distance between two probabilities $P=p\cdot\mu$ and $R=r\cdot\mu$ dominated by a measure μ on a measurable space $(E,\mathcal{E})$ is given by

$$h(P,R)=\left[\frac{1}{2}\int_E\left(\sqrt{p}-\sqrt{r}\right)^2 d\mu\right]^{1/2},$$

the result being independent of the choice of the dominating measure μ.

On the basis of the observations $X_1,\dots,X_n$, we build an estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}^\star$ with values in $\overline{\boldsymbol{\Theta}}$ and evaluate its performance by the quantity $\mathbf{h}^2(\mathbf{Q}^\star,\mathbf{Q}_{\widehat{\boldsymbol{\theta}}})$ with the notations

$$\mathbf{Q}^\star=(Q_1^\star,\dots,Q_n^\star)\quad\text{and}\quad\mathbf{Q}_{\boldsymbol{\theta}}=(Q_{\boldsymbol{\theta}},\dots,Q_{\boldsymbol{\theta}})$$

where $\boldsymbol{\theta}$ denotes a function from $\mathscr{W}$ into I.

Our aim is to design $\widehat{\boldsymbol{\theta}}$ not only to guarantee that $\mathbf{h}^2(\mathbf{Q}^\star,\mathbf{Q}_{\widehat{\boldsymbol{\theta}}})$ is small but also that this quantity remains stable to a possible departure from the assumptions we started from i.e. when $\mathbf{Q}^\star\notin\{\mathbf{Q}_{\boldsymbol{\theta}},\ \boldsymbol{\theta}\in\overline{\boldsymbol{\Theta}}\}$ but $\inf_{\boldsymbol{\theta}\in\overline{\boldsymbol{\Theta}}}\mathbf{h}^2(\mathbf{Q}^\star,\mathbf{Q}_{\boldsymbol{\theta}})$ is small.

If P_i and P_i' denote two distributions for a random variable (W_i,Y_i) the conditional distributions of which given W_i are Q_i and Q_i' respectively, then

$$h^2(P_i,P_i')=\int_{\mathscr{W}}h^2\left(Q_i(w),Q_i'(w)\right)dP_{W_i}(w)$$

and we shall write $P_i=Q_i\cdot P_{W_i}$ and $P_i'=Q_i'\cdot P_{W_i}$. In particular, for all functions $\boldsymbol{\theta}:\mathscr{W}\to I$

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

where $P_i^\star=Q_i^\star\cdot P_{W_i}$ and $P_{i,\boldsymbol{\theta}}=Q_{\boldsymbol{\theta}}\cdot P_{W_i}$ for all $i\in\{1,\dots,n\}$. The probability $\mathbf{P}^\star=\otimes_{i=1}^n P_i^\star$ corresponds to the true distribution of the observed data $\boldsymbol{X}=(X_1,\dots,X_n)$ while $\mathbf{P}_{\boldsymbol{\theta}}=\otimes_{i=1}^n P_{i,\boldsymbol{\theta}}$ denotes 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 is given by $Q_{\boldsymbol{\theta}(W_i)}\in\overline{\mathscr{Q}}$ for all i. Unlike $\mathbf{Q}_{\widehat{\boldsymbol{\theta}}}$, $\mathbf{P}_{\widehat{\boldsymbol{\theta}}}$ is not an estimator (of $\mathbf{P}^\star$) since the distributions of the W_i are unknown. Nevertheless it will sometimes be convenient to interpret our results in terms of an Hellinger-type distance between $\mathbf{P}^\star$ and $\mathbf{P}_{\widehat{\boldsymbol{\theta}}}$. For these reasons we shall sometimes write $\mathbf{h}(\mathbf{P}^\star,\mathbf{P}_{\widehat{\boldsymbol{\theta}}})$ for $\mathbf{h}(\mathbf{Q}^\star,\mathbf{Q}_{\widehat{\boldsymbol{\theta}}})$ and $\mathbf{h}(\mathbf{P}_{\boldsymbol{\theta}},\mathbf{P}_{\boldsymbol{\theta}'})$ for $\mathbf{h}(\mathbf{Q}_{\boldsymbol{\theta}},\mathbf{Q}_{\boldsymbol{\theta}'})$ when $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ denote mappings from $\mathscr{W}$ into I.

2.1. **Examples.** Let us present here some typical statistical settings to which our approach applies.

Example 2 (Gaussian regression with known variance)**.** Given n independent random variables $W_1,\dots,W_n$ with values in $\mathscr{W}$, let

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

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) with $A(\theta) = \theta^2/(2\sigma^2)$ and $S(y) = y/\sigma^2$ for all $y \in \mathbb{R}$.

Example 3 (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

$$\mathbb{P}\left[Y_i = y | W_i\right] = \frac{\exp\left[y\boldsymbol{\theta}^\star(W_i)\right]}{1 + \exp\left[\boldsymbol{\theta}^\star(W_i)\right]} \quad \text{for all } y \in \{0, 1\}. \tag{4}$$

This means that the conditional distribution of Y_i given W_i is Bernoulli with mean $\left(1 + \exp\left[-\boldsymbol{\theta}^\star(W_i)\right]\right)^{-1}$ for some regression function $\boldsymbol{\theta}^\star$ with values in $I = \mathbb{R}$. This model is equivalent to the logit one presented in Example 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 thus a density with respect to the counting measure ν on $\mathscr{Y} = \{0, 1\}$ of the form (2) with $A(\theta) = \log(1+e^\theta)$ and $s(y) = y$ for all $y \in \mathscr{Y}$.

Example 4 (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) 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 is presumed to be Poisson with mean $\exp\left[\boldsymbol{\theta}^\star(W_i)\right]$ for some regression function $\boldsymbol{\theta}^\star$ with values in $I = \mathbb{R}$.

Example 5 (Exponential multiplicative regression)**.** The random variables $W_1, \dots, W_n$ are independent and

$$Y_i = \frac{Z_i}{\boldsymbol{\theta}^\star(W_i)} \quad \text{for all } i \in \{1, \dots, n\} \tag{5}$$

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 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) with $S(y) = -y$ for all $y \in \mathscr{Y} = \mathbb{R}_+$ and $A(\theta) = -\log\theta$.

3. The main results

3.1. **The estimation procedure.** As mentioned in the introduction, our approach is based on ρ-estimation. We shall not recall here the basic ideas that underline the construction of these estimators and rather refer the

reader to Baraud and Birgé (2018). Let ψ be the function defined on $[0,+\infty]$ by

$$\psi(x)=\frac{x-1}{x+1} \quad \text{for } x\in[0,+\infty) \quad \text{and} \quad \psi(+\infty)=1. \tag{6}$$

In order to avoid measurability issues, we restrict ourselves to a finite or countable subset $\boldsymbol{\Theta}$ of $\overline{\boldsymbol{\Theta}}$ and define

$$\mathbf{T}(\boldsymbol{X},\boldsymbol{\theta},\boldsymbol{\theta}')=\sum_{i=1}^{n}\psi\left(\sqrt{\frac{q_{\boldsymbol{\theta}'}(X_i)}{q_{\boldsymbol{\theta}}(X_i)}}\right) \quad \text{for} \quad \boldsymbol{\theta},\boldsymbol{\theta}'\in\boldsymbol{\Theta}, \tag{7}$$

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

$$\boldsymbol{\upsilon}(\boldsymbol{X},\boldsymbol{\theta})=\sup_{\boldsymbol{\theta}'\in\boldsymbol{\Theta}}\mathbf{T}(\boldsymbol{X},\boldsymbol{\theta},\boldsymbol{\theta}') \quad \text{for all} \quad \boldsymbol{\theta}\in\boldsymbol{\Theta} \tag{8}$$

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

$$\mathscr{E}(\boldsymbol{X})=\left\{\boldsymbol{\theta}\in\boldsymbol{\Theta} \text{ such that } \boldsymbol{\upsilon}(\boldsymbol{X},\boldsymbol{\theta})\leqslant\inf_{\boldsymbol{\theta}'\in\boldsymbol{\Theta}}\boldsymbol{\upsilon}(\boldsymbol{X},\boldsymbol{\theta}')+1\right\}. \tag{9}$$

The random variable $\widehat{\boldsymbol{\theta}}(\boldsymbol{X})$ is our estimator of the regression function $\boldsymbol{\theta}^\star$ and we recall that $\mathbf{Q}_{\widehat{\boldsymbol{\theta}}}=(Q_{\widehat{\boldsymbol{\theta}}},\ldots,Q_{\widehat{\boldsymbol{\theta}}})$.

The construction of the estimator is only based on the choices of the exponential family given by (2) 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.

In the right-hand side of (9), the additive constant 1 plays no magic role and can be replaced by any smaller positive number. Whenever possible, the choice of an estimator $\widehat{\boldsymbol{\theta}}$ satisfying $\boldsymbol{\upsilon}(\boldsymbol{X},\widehat{\boldsymbol{\theta}})=\inf_{\boldsymbol{\theta}\in\overline{\boldsymbol{\Theta}}}\boldsymbol{\upsilon}(\boldsymbol{X},\boldsymbol{\theta})$ should be preferred.

The fact that we build our estimator on a finite or countable subset $\boldsymbol{\Theta}$ of $\overline{\boldsymbol{\Theta}}$ will not be 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.

3.2. **The main assumption.** Let us make the following assumption:

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

We recall that $\overline{\boldsymbol{\Theta}}$ is VC-subgraph on $\mathscr{W}$ with dimension not larger than $V\geqslant 1$ if, whatever the finite subset $\mathcal{S}$ of $V+1$ points in $\mathscr{W}\times\mathbb{R}$, there exists at least one subset S of $\mathcal{S}$ that is not the intersection of $\mathcal{S}$ with an (open) subgraph of a function in $\overline{\boldsymbol{\Theta}}$, i.e.

$$S\neq\mathcal{S}\cap\{(w,t)\in\mathscr{W}\times\mathbb{R},\ \boldsymbol{\theta}(w)>t\} \quad \text{whatever } \boldsymbol{\theta}\in\overline{\boldsymbol{\Theta}}. \tag{10}$$

Assumption 1 is satisfied when $\overline{\boldsymbol{\Theta}}$ is a linear space $\mathcal{V}$ with finite dimension $d \geqslant 1$, in which case $V = d + 1$. It also holds for any set $\overline{\boldsymbol{\Theta}}$ of the form $\{F(\boldsymbol{\beta}),\ \boldsymbol{\beta} \in \mathcal{V}\}$ where F is a monotone function on the real-line.

Another situation, although elementary, is that of a finite set $\overline{\boldsymbol{\Theta}}$. As soon as the cardinality k of a set $\mathcal{S}$ satisfies $2^k > \operatorname{Card} \overline{\boldsymbol{\Theta}}$, there exists $S \subset \mathcal{S}$ that fulfils (10). This means that $\overline{\boldsymbol{\Theta}}$ is VC-subgraph and its dimension is not larger than $V = (\log(\operatorname{Card} \overline{\boldsymbol{\Theta}})/\log 2) \vee 1$.

For more properties of the classes of functions which are VC-subgraph, we refer the reader to van der Vaart and Wellner (1996)[Section 2.6.2].

3.3. **The performance of the estimator $\widehat{\boldsymbol{\theta}}$.** Let us set

$$c_1 = 150, \quad c_2 = 1.1 \times 10^6, \quad c_3 = 5014 \tag{11}$$

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

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

The risk of our estimator satisfies the following properties.

Theorem 1. *Let $\xi > 0$. Under Assumption 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 $\widehat{\boldsymbol{\theta}}$ defined in Section 3.1 satisfies, with a probability at least $1 - e^{-\xi}$,*

$$\mathbf{h}^2\left(\mathbf{Q}^\star, \mathbf{Q}_{\widehat{\boldsymbol{\theta}}}\right) \leqslant c_1 \mathbf{h}^2(\mathbf{Q}^\star, \boldsymbol{\mathcal{Q}}) + c_2 V \left[9.11 + \log_+\left(\frac{n}{V}\right)\right] + c_3\left(1.5 + \xi\right) \tag{12}$$

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

If, in particular, $\boldsymbol{\mathcal{Q}}$ is dense in $\overline{\boldsymbol{\mathcal{Q}}} = \{\mathbf{Q}_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}}\}$ with respect to the Hellinger-type distance $\mathbf{h}$, there exists a numerical constant $C > 0$ such that the estimator $\widehat{\boldsymbol{\theta}}$ satisfies

$$C\mathbf{h}^2\left(\mathbf{Q}^\star, \mathbf{Q}_{\widehat{\boldsymbol{\theta}}}\right) \leqslant \mathbf{h}^2(\mathbf{Q}^\star, \overline{\boldsymbol{\mathcal{Q}}}) + V\left[1 + \log_+\left(\frac{n}{V}\right)\right] + \xi \tag{13}$$

with a probability at least $1 - e^{-\xi}$.

The difference between (12) and (13) depends on the approximation properties of $\boldsymbol{\mathcal{Q}} = \{\mathbf{Q}_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ with respect to $\overline{\boldsymbol{\mathcal{Q}}} = \{\mathbf{Q}_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}}\}$ (for the Hellinger-type distance $\mathbf{h}$). Actually, whatever $\mathbf{Q}^\star$,

$$\left|\mathbf{h}(\mathbf{Q}^\star, \boldsymbol{\mathcal{Q}}) - \mathbf{h}(\mathbf{Q}^\star, \overline{\boldsymbol{\mathcal{Q}}})\right| \leqslant \sup_{\overline{\boldsymbol{\theta}} \in \overline{\boldsymbol{\Theta}}} \mathbf{h}\left(\mathbf{Q}_{\overline{\boldsymbol{\theta}}}, \boldsymbol{\mathcal{Q}}\right) \tag{14}$$

and, if the right-hand side is not larger than some $\eta > 0$, i.e. if $\boldsymbol{\mathcal{Q}}$ is an η-net for $\overline{\boldsymbol{\mathcal{Q}}}$, (12) leads to

$$C\mathbf{h}^2\left(\mathbf{Q}^\star, \mathbf{Q}_{\widehat{\boldsymbol{\theta}}}\right) \leqslant \mathbf{h}^2(\mathbf{Q}^\star, \overline{\boldsymbol{\mathcal{Q}}}) + \eta^2 + V\left[1 + \log_+(n/V)\right] + \xi$$

for some suitable numerical constant $C > 0$.

When $\boldsymbol{\Theta}$ is dense in $\overline{\boldsymbol{\Theta}}$ for the topology of pointwise convergence, it can be shown that one can always take $\eta = 0$. We shall 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 shall assume for the sake of simplicity that $\eta = 0$, 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.

In order to comment further on Theorem 1, we shall present (13) in a slightly different form. We have seen in Section 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 (13) is therefore equivalent to

$$C\mathbf{h}^2\left(\mathbf{P}^\star, \mathbf{P}_{\widehat{\boldsymbol{\theta}}}\right) \leqslant \mathbf{h}^2(\mathbf{P}^\star, \overline{\mathscr{P}}) + V\left[1 + \log_+\left(\frac{n}{V}\right)\right] + \xi \tag{15}$$

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 $\widehat{\boldsymbol{\theta}}$

$$C\mathbb{E}\left[\mathbf{h}^2\left(\mathbf{P}^\star, \mathbf{P}_{\widehat{\boldsymbol{\theta}}}\right)\right] \leqslant \mathbf{h}^2(\mathbf{P}^\star, \overline{\mathscr{P}}) + V\left[1 + \log_+\left(\frac{n}{V}\right)\right]. \tag{16}$$

In order to comment upon (16), let us start with the ideal situation where $\mathbf{P}^\star$ belongs to $\overline{\mathscr{P}}$, i.e. $\mathbf{P}^\star = \mathbf{P}_{\boldsymbol{\theta}^\star}$ with $\boldsymbol{\theta}^\star \in \overline{\boldsymbol{\Theta}}$, in which case (16) leads to

$$C\mathbb{E}\left[\mathbf{h}^2\left(\mathbf{P}_{\boldsymbol{\theta}^\star}, \mathbf{P}_{\widehat{\boldsymbol{\theta}}}\right)\right] \leqslant V\left[1 + \log_+\left(\frac{n}{V}\right)\right]. \tag{17}$$

Up to the logarithmic factor, the right-hand side of this inequality is of the expected order of magnitude V for the quantity $\mathbf{h}^2(\mathbf{P}_{\boldsymbol{\theta}^\star}, \mathbf{P}_{\widehat{\boldsymbol{\theta}}})$.

When the true distribution $\mathbf{P}^\star$ writes as $\mathbf{P}_{\boldsymbol{\theta}^\star}$ but the regression function $\boldsymbol{\theta}^\star$ does not belong to $\overline{\boldsymbol{\Theta}}$, or if the conditional distributions of the Y_i given W_i do not belong to our exponential family, inequality (16) shows that, as compared to (17), 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 (17) 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 our data set has been contaminated in such a way that the true distribution $\mathbf{P}^\star$ of $\boldsymbol{X} = (X_1, \ldots, X_n)$ writes as $\left[(1-\alpha)P_{\overline{\boldsymbol{\theta}}} + \alpha R\right]^{\otimes n}$ with $\alpha \in (0, 1/2)$, $\overline{\boldsymbol{\theta}} \in \overline{\boldsymbol{\Theta}}$ and an arbitrary distribution $R \neq P_{\overline{\boldsymbol{\theta}}}$ or, alternatively, that the data set contains a proportion $k/n \leqslant \alpha$ of outliers $a_1, \ldots, a_k \in \mathscr{X}$, in which case we may write $\mathbf{P}^\star$ as $\bigotimes_{i=1}^{n-k} Q_{\overline{\boldsymbol{\theta}}} \bigotimes_{i=1}^{k} \delta_{a_i}$. Using the classical inequality $h^2 \leqslant D$ where D denotes

the total variation distance between probabilities, we get

$$\mathbf{h}^2(\mathbf{P}^\star, \overline{\mathscr{P}}) \leqslant \mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\overline{\boldsymbol{\theta}}}) \leqslant \sum_{i=1}^{n} D(P_i^\star, P_{\overline{\boldsymbol{\theta}}}) \leqslant n\alpha \tag{18}$$

which means that whenever $n\alpha$ 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.

3.4. **From a natural to a general exponential family.** In Section 2, 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

$$r_\gamma : y \mapsto e^{u(\gamma)S(y) - B(\gamma)} \quad \text{for } \gamma \in J. \tag{19}$$

In (19), J denotes a (non-trivial) 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 $\overline{\boldsymbol{\mathscr{Q}}}$ as

$$\overline{\boldsymbol{\mathscr{Q}}} = \left\{\mathbf{R}_{\boldsymbol{\gamma}} = (R_{\boldsymbol{\gamma}}, \ldots, R_{\boldsymbol{\gamma}})\ \boldsymbol{\gamma} \in \overline{\boldsymbol{\Gamma}}\right\} \tag{20}$$

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 could 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 3.1 under the natural parametrization of the exponential family $\overline{\mathscr{Q}}$ and the finite or countable model $\boldsymbol{\Theta} = \{\boldsymbol{\theta} = u \circ \boldsymbol{\gamma},\ \boldsymbol{\gamma} \in \boldsymbol{\Gamma}\}$.

Since our model $\overline{\boldsymbol{\mathscr{Q}}}$ for the conditional probabilities $\mathbf{Q}^\star$ is unchanged (only its parametrization 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 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 1 applies. The following corollary is therefore straightforward.

Corollary 1. *Let $\xi > 0$. If the statistical model $\overline{\boldsymbol{\mathscr{Q}}}$ is under the general form (20) and $\overline{\boldsymbol{\Gamma}}$ is VC-subgraph with dimension not larger than $V \geqslant 1$, 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 $\widehat{\boldsymbol{\gamma}}$ satisfies with a probability*

at least $1-e^{-\xi}$,

$$\mathbf{h}^2\left(\mathbf{Q}^\star, \mathbf{R}_{\widehat{\gamma}}\right) \leqslant c_1 \mathbf{h}^2(\mathbf{Q}^\star, \boldsymbol{\mathscr{Q}}) + c_2 V\left[9.11 + \log_+\left(\frac{n}{V}\right)\right] + c_3\left(1.5+\xi\right) \tag{21}$$

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

$$\mathbb{E}\left[\mathbf{h}^2\left(\mathbf{Q}^\star, \mathbf{R}_{\widehat{\gamma}}\right)\right] \leqslant C'\left[\mathbf{h}^2(\mathbf{Q}^\star, \boldsymbol{\mathscr{Q}}) + V\left[1 + \log_+\left(\frac{n}{V}\right)\right]\right], \tag{22}$$

for some numerical constant $C' > 0$.

A nice feature of our approach lies in the fact that (21) holds for all exponential families simultaneously and all ways of parametrizing them.

4. Uniform risk bounds

Throughout this section, we shall assume that the W_i are i.i.d. with common distribution P_W and that $\mathbf{Q}^\star = \mathbf{R}_{\boldsymbol{\gamma}^\star}$ belongs to a statistical model of the (general) form given by (20) where $\overline{\boldsymbol{\Gamma}}$ is a class $\mathcal{H}$ of smooth functions. Our aim is to estimate the regression function $\boldsymbol{\gamma}^\star$ under the assumption that it belongs to $\mathcal{H}$. More precisely, we want to evaluate the minimax risk over $\mathcal{H}$, i.e. the quantity

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

where the infimum runs among all estimators $\widetilde{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}^\star$ based on the n-sample $X_1, \ldots, X_n$ and, for all functions $\boldsymbol{\gamma}, \boldsymbol{\gamma}'$ from $\mathscr{W}$ into J,

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

We have seen that Corollary 1 shows that our risk bound is invariant under a parameter change $\boldsymbol{\gamma} = v(\boldsymbol{\theta})$, with $v = u^{-1}$. This property is due to the fact that the composition by v preserves the VC-subgraph property and the VC-dimension. The situation changes dramatically when $\boldsymbol{\theta}$ is assumed to belong to a smoothness class since a parameter change does not preserve smoothness in general. This means that, under smoothness assumptions, the order of magnitude of $\mathcal{R}_n(\mathcal{H})$ with respect to the sample size n depends on our way of parametrizing the exponential family $\overline{\mathscr{Q}}$ or equivalently of choosing v.

What we want to do in this section is to establish upper and lower bounds on the quantity $\mathcal{R}_n(\mathcal{H})$ that hold true for all choices of exponential families (considered as families of probabilities) simultaneously but under some specific choices of their parametrizations. Let us first investigate the situation where the exponential family is parametrized by the mean.

4.1. **Parametrizing by the mean.** A common parametrization of an exponential family $\overline{\mathscr{Q}} = \{R_\gamma,\ \gamma \in J\}$ is by the means of the distributions, i.e. $\gamma = \int_{\mathscr{Y}} y dR_\gamma(y)$. This is typically the case for the Bernoulli, Gaussian and Poisson families for example. Our observations $(W_1, Y_1), \ldots, (W_n, Y_n)$ then satisfy the equality

$$Y_i = \boldsymbol{\gamma}^\star(W_i) + \varepsilon_i \quad \text{with} \quad \mathbb{E}\left[\varepsilon_i | W_i\right] = 0 \quad \text{for all } i \in \{1, \ldots, n\}. \tag{23}$$

With the relationship between Y_i and W_i written in this form, the problem becomes equivalent to that of estimating a regression function in a regression model and one might expect that the rates for estimating $\boldsymbol{\gamma}^\star$ under smoothness assumptions be similar to those established when the errors are assumed to be i.i.d. Gaussian random variables. Unfortunately, this is not true in general since the regression model given by (23) is actually heteroscedastic and the variances of the errors depend on the values of the regression function. We shall now show that, in this case, the rates can be much slower than those we would get in the Gaussian case.

For $\alpha \in (0, 1]$ and $M > 0$, let $\mathcal{H} = \mathcal{H}_\alpha(M)$ be the set of functions $\boldsymbol{\gamma}$ on $[0, 1]$ with values in J that satisfy the Hölder condition

$$|\boldsymbol{\gamma}(x) - \boldsymbol{\gamma}(y)| \leqslant M|x - y|^\alpha \quad \text{for all } x, y \in [0, 1]. \tag{24}$$

We prove

Proposition 1. *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 $n \geqslant 1$,*

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

In the Poisson case, 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 in the Gaussian case, namely $n^{-2\alpha/(2\alpha+1)}$. We conclude that the parametrization by the mean leads to minimax rates that do depend on the exponential family. In order to obtain a rate that is independent of the exponential family, we must turn to another way of parametrizing them.

4.2. **Connecting the Hellinger distance to the $\mathbb{L}_2$-norm.** The unusual lower bound established in Proposition 1 can actually be explained as follows. When the Poisson family $\overline{\mathscr{Q}}$ 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 $h^2(R_\gamma, R_{\gamma'})$ writes as

$$h^2(R_\gamma, R_{\gamma'}) = \int_{\mathscr{W}} \left[1 - e^{-\left(\sqrt{\gamma(w)} - \sqrt{\gamma'(w)}\right)^2/2} \right] dP_W(w) \tag{25}$$

and therefore behaves like

$$\frac{1}{2}\left\|\sqrt{\boldsymbol{\gamma}}-\sqrt{\boldsymbol{\gamma}'}\right\|_2^2=\frac{1}{2}\int_{\mathscr{W}}\left(\sqrt{\boldsymbol{\gamma}}-\sqrt{\boldsymbol{\gamma}'}\right)^2 dP_W$$

whenever $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$ are close for the supremum norm.

In contrast, if $\overline{\mathscr{Q}}$ is the family of Gaussian distributions with variance 1, we obtain that

$$h^2(R_{\boldsymbol{\gamma}},R_{\boldsymbol{\gamma}'})=\int_{\mathscr{W}}\left[1-e^{-(\boldsymbol{\gamma}(w)-\boldsymbol{\gamma}'(w))^2/8}\right]dP_W(w) \tag{26}$$

and this quantity behaves like $\|\boldsymbol{\gamma}-\boldsymbol{\gamma}'\|_2^2/8$ whenever $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$ are close for the supremum norm.

Unless one assumes that the functions $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$ are bounded away from 0 and infinity, the losses $h^2(R_{\boldsymbol{\gamma}},R_{\boldsymbol{\gamma}'})$ are not equivalent in the Poisson and Gaussian cases. In fact, the risk for estimating a regression function $\boldsymbol{\gamma}^\star\in\mathcal{H}=\mathcal{H}_\alpha(M)$ in the Poisson case is expected to be of the same order as that for estimating $\sqrt{\boldsymbol{\gamma}^\star}$ in the Gaussian case. If $\boldsymbol{\gamma}^\star$ is assumed to be of regularity α, that of $\sqrt{\boldsymbol{\gamma}^\star}$ is in general not better than $\alpha/2$. The lower bound we established in Proposition 1 actually corresponds to the usual (Gaussian) rate for estimating functions with regularity $\alpha/2$.

In order to avoid this phenomenon, we need to choose a parametrization of $\overline{\mathscr{Q}}=\{R_\gamma,\ \gamma\in J\}$ that makes the quantities $h^2(R_{\boldsymbol{\gamma}},R_{\boldsymbol{\gamma}'})$ equivalent in all exponential families, at least when $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$ are close in supremum norm. To achieve this goal it is actually enough to make all these quantities locally equivalent to a fixed one and we shall choose the $\mathbb{L}_2(P_W)$-norm between the functions $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$. For example, in the Poisson case, a look at (25) shows that taking for R_γ the Poisson distribution with parameter γ^2 instead of γ would make this connection possible.

For a general exponential family $\overline{\mathscr{Q}}=\{Q_\theta,\ \theta\in I\}$ given by (2) under its natural form, we shall look for a parametrization $\gamma=v(\theta)$ for which $h^2(Q_\theta,Q_{\theta'})=h^2(R_\gamma,R_{\gamma'})$ is of order $|\gamma-\gamma'|^2$ whenever γ and γ' are close enough. When I is an open interval, it is well known from Ibragimov and Has'minskiĭ (1981) that the Hellinger distance $h^2(R_\gamma,R_{\gamma'})$ behaves locally like $\mathcal{I}(\gamma)|\gamma-\gamma'|^2/8$ where $\mathcal{I}(\gamma)$ denotes the Fisher information at $\gamma=v(\theta)$, i.e. the quantity given by the formula $A''(\theta)[v'(\theta)]^{-2}$. As a consequence, if we want that $h(R_\gamma,R_{\gamma'})$ be locally equivalent to $|\gamma-\gamma'|$ up to a multiplicative constant that do not depend on γ, it suffices to choose the parametrization v in such a way that this Fisher information is constant, say equal to 8. This means that v must satisfy

$$v'(\theta)=\sqrt{\frac{A''(\theta)}{8}}>0\quad\text{for all }\theta\in I. \tag{27}$$

Since the function A'' is continuous and positive on I, the function v satisfying (27) is increasing and defines a $\mathscr{C}^1$-diffeomorphism from I onto an

open interval $J = v(I)$, which means that v and $u = v^{-1}$ are both continuously differentiable on I and J respectively. The exponential family $\overline{\mathscr{Q}}$ remains thus regular under the parametrization $\gamma = v(\theta)$ and we can derive the following results.

Proposition 2. *Let $v = u^{-1}$ be a function that satisfies (27) on the open interval I. If the exponential family $\overline{\mathscr{Q}}$ is parametrized by $\gamma = v(\theta)$, i.e. $\overline{\mathscr{Q}} = \{R_\gamma = Q_{u(\gamma)},\ \gamma \in J\}$ then for all functions $\boldsymbol{\gamma}, \boldsymbol{\gamma}'$ on $\mathscr{W}$ with values in J,*

$$h^2\left(R_{\boldsymbol{\gamma}}, R_{\boldsymbol{\gamma}'}\right) \leqslant \left\|\boldsymbol{\gamma} - \boldsymbol{\gamma}'\right\|_2^2 = \int_{\mathscr{W}} \left(\boldsymbol{\gamma} - \boldsymbol{\gamma}'\right)^2 dP_W. \tag{28}$$

Besides, for all compact subset K of J, there exists a constant $c_K > 0$ such that for all functions $\boldsymbol{\gamma}, \boldsymbol{\gamma}'$ on $\mathscr{W}$ with values in K,

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

This result implies that, under this new parametrization, the Hellinger-type distance h between $R_{\boldsymbol{\gamma}}$ and $R_{\boldsymbol{\gamma}'}$ is equivalent to the $\mathbb{L}_2(P_W)$-distance between the functions $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$, at least when $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$ take their values in a compact subset of J.

It is well-known that the functions $v_1 : \theta \mapsto (1/\sqrt{2})\arcsin(1/\sqrt{1+e^{-\theta}})$, $v_2 : \theta \mapsto (1/\sqrt{2})e^{\theta/2}$ for $\theta \in \mathbb{R}$ and $v_3 : \theta \mapsto (\sqrt{8})^{-1}\log\theta$ for $\theta > 0$ all satisfy Condition (27) in the cases of Examples 3, 4 and 5 respectively.

4.3. **Uniform risk bounds over Hölder classes.** Let us now assume that the exponential family $\overline{\mathscr{Q}} = \{R_\gamma,\ \gamma \in J\}$ has been parametrized in such a way that there exists a constant $\kappa > 0$ such that

$$h(R_\gamma, R_{\gamma'}) \leqslant \kappa\left|\gamma - \gamma'\right| \quad \text{for all } \gamma, \gamma' \in J \tag{29}$$

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

$$h(R_\gamma, R_{\gamma'}) \geqslant c_K\left|\gamma - \gamma'\right| \quad \text{for all } \gamma, \gamma' \in K. \tag{30}$$

We have seen in Proposition 2 that (29) and (30) are both satisfied (with $\kappa = 1$) as soon as γ is a function v of the natural parameter θ that fulfils (27).

For $\alpha \in (0, 1]$ and $M > 0$, we consider the class $\mathcal{H}_\alpha(M)$ of functions $\boldsymbol{\gamma}$ defined on $\mathscr{W} = [0, 1]$ with values in J that satisfy (24). We can then derive the following upper bound for the risk. It holds whatever the distributions P_{W_i}.

Proposition 3. *Let $\alpha \in (0, 1]$ and $M > 0$. If (29) is satisfied, then*

$$\mathcal{R}_n\left(\mathcal{H}_\alpha(M)\right) \leqslant 2C'\left[\left(\frac{(\kappa M)^{1/\alpha}\log(en)}{n}\right)^{\frac{2\alpha}{1+2\alpha}} + \frac{3\log(en)}{2n}\right] \tag{31}$$

where C' is the numerical constant appearing in (22).

The next result shows that, up to a logarithmic factor, the rate $n^{-2\alpha/(1+2\alpha)}$ is optimal, at least when the W_i are uniformly distributed on $[0,1]$.

Proposition 4. *Let $\alpha \in (0,1]$, $M > 0$ and $n \geqslant 1$. If P_W is the uniform distribution on $[0,1]$ and the inequalities (29) and (30) are satisfied for an interval K of length $2\overline{L} > 0$*

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

It follows from Proposition 2 that the risk bound (31) holds with $\kappa = 1$ when the exponential family is parametrized by $\gamma = v(\theta)$ and v satisfies (27). Proposition 3 shows that it is possible to establish a risk bound of optimal order on $\mathcal{R}_n(\mathcal{H}_\alpha(M))$ in great generality i.e. independently of the exponential family, provided that it is suitably parametrized.

The result established in Proposition 3 relies on the crucial property that ρ-estimators still perform correctly on a parameter space that does not necessarily contain the true parameter but only provides a suitable approximation of it. We may therefore use (21) and the well-known fact that the Hölder class $\mathcal{H}_\alpha(M)$ can be well approximated in supremum norm, hence in $\mathbb{L}_2(P_W)$-norm, by linear spaces which are VC-classes and (28) enables us to control the Hellinger-type approximation of these VC-classes by the $\mathbb{L}_2(P_W)$-one. Our approach does not restricted to Hölder classes and can easily be extended to any functional spaces $\mathcal{H}$ the elements of which are well approximated by linear spaces.

5. Calculation of ρ-estimators and simulation study

Let us now go back to Examples 3, 4, 5 which correspond respectively to the Bernoulli, Poisson and exponential distributions parametrized by their natural parameters. For this three cases, we shall illustrate by a simulation study the respective performance of the ρ-estimator $\widehat{\boldsymbol{\theta}}$ and the MLE in two different situations: when the statistical model is exact and when it is not. For the Poisson and exponential distributions, we shall in addition study a median-based estimator $\widehat{\boldsymbol{\theta}}_0$ which is defined as any minimizer over $\overline{\boldsymbol{\Theta}}$ of the criterion

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

where $m(\theta)$ is the median of the distribution Q_θ, $\theta \in I$, or an approximation of it. We set $m(\theta) = e^\theta + 1/3 - 0.02e^{-\theta}$ for the Poisson distribution and $m(\theta) = (\log 2)/\theta$ for the exponential one. This estimator is chosen for its robustness properties with respect to contamination and outliers.

Throughout this study, $\mathscr{W} = \mathbb{R}^5$ and our model $\overline{\boldsymbol{\Theta}}$ for the regression function $\boldsymbol{\theta}^\star$ is the set of all functions $\boldsymbol{\theta}$ on $\mathscr{W}$ with values in I, where $I = \mathbb{R}$

in the logit and Poisson models and $I = (0, +\infty)$ in the exponential one, that take the following forms for each of these exponential families: for all $w = (w_1, \ldots, w_5) \in \mathscr{W}$

$$\boldsymbol{\theta}(w) = \begin{cases} \eta_0 + \sum_{j=1}^5 \eta_j w_j & \text{(logit)} \\ \log\log\left[1 + \exp\left(\eta_0 + \sum_{j=1}^5 \eta_j w_j\right)\right] & \text{(Poisson)} \\ \log\left[1 + \exp\left(\eta_0 + \sum_{j=1}^5 \eta_j w_j\right)\right] & \text{(exponential)} \end{cases} \tag{32}$$

where $\eta = (\eta_0, \ldots, \eta_5)$ belongs to $\mathbb{R}^6$. In particular, when $\mathbf{P}^\star = \mathbf{P}_{\boldsymbol{\theta}}$ with $\boldsymbol{\theta}$ is given by (32), the conditional expectation of Y_1 given $W_1 = w$ satisfies

$$\mathbb{E}\left[Y_1 | W_1 = w\right] = \begin{cases} \log\left[1 + \exp\left(\eta_0 + \sum_{j=1}^5 \eta_j w_j\right)\right] & \text{(Poisson)} \\ \left[\log\left[1 + \exp\left(\eta_0 + \sum_{j=1}^5 \eta_j w_j\right)\right]\right]^{-1} & \text{(exponential).} \end{cases}$$

The set $\overline{\boldsymbol{\Theta}}$ is VC-subgraph with dimension not larger than 7.

5.1. **Calculation of the ρ-estimator.** As mentioned in Section 3, we call ρ-estimator $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\boldsymbol{X})$ any element of the random set

$$\mathscr{E}(\boldsymbol{X}) = \left\{ \boldsymbol{\theta} \in \boldsymbol{\Theta} \text{ such that } \boldsymbol{\upsilon}(\boldsymbol{X}, \boldsymbol{\theta}) \leqslant \inf_{\boldsymbol{\theta}' \in \boldsymbol{\Theta}} \boldsymbol{\upsilon}(\boldsymbol{X}, \boldsymbol{\theta}') + 1 \right\},$$

where

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

An interesting feature of this definition lies in the fact that if $\widehat{\boldsymbol{\theta}}$ satisfies $\boldsymbol{\upsilon}(\boldsymbol{X}, \widehat{\boldsymbol{\theta}}) \leqslant \varepsilon \leqslant 1$ it is necessarily a ρ-estimator. Indeed, since $\boldsymbol{\upsilon}(\boldsymbol{X}, \boldsymbol{\theta}') \geqslant \mathbf{T}(\boldsymbol{X}, \boldsymbol{\theta}', \boldsymbol{\theta}') = 0$ for all $\boldsymbol{\theta}' \in \boldsymbol{\Theta}$, if $\boldsymbol{\upsilon}(\boldsymbol{X}, \widehat{\boldsymbol{\theta}}) \leqslant \varepsilon$ then

$$\boldsymbol{\upsilon}(\boldsymbol{X}, \widehat{\boldsymbol{\theta}}) \leqslant \varepsilon \leqslant \inf_{\boldsymbol{\theta}' \in \boldsymbol{\Theta}} \boldsymbol{\upsilon}(\boldsymbol{X}, \boldsymbol{\theta}') + \varepsilon \leqslant \inf_{\boldsymbol{\theta}' \in \boldsymbol{\Theta}} \boldsymbol{\upsilon}(\boldsymbol{X}, \boldsymbol{\theta}') + 1,$$

and $\widehat{\boldsymbol{\theta}}$ therefore belongs to $\mathscr{E}(\boldsymbol{X})$. Consequently, in order to calculate our ρ-estimator, we may look for an element $\widehat{\boldsymbol{\theta}}$ that satisfies $\boldsymbol{\upsilon}(\boldsymbol{X}, \widehat{\boldsymbol{\theta}}) \leqslant \varepsilon$ and, to find such a $\widehat{\boldsymbol{\theta}}$, we use the following algorithm. We first initialise the process at some value $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$. We know from Baraud and Birgé (2018) that, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the quantity $\mathbf{T}(\boldsymbol{X}, \boldsymbol{\theta}_0, \boldsymbol{\theta})$ is an estimator of the difference $\mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_{\boldsymbol{\theta}_0}) - \mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_{\boldsymbol{\theta}})$. More precisely, the following inequalities hold:

$$\mathbb{E}\left[\mathbf{T}(\boldsymbol{X}, \boldsymbol{\theta}_0, \boldsymbol{\theta})\right] \begin{cases} \leqslant a_0 \mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_{\boldsymbol{\theta}_0}) - a_1 \mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_{\boldsymbol{\theta}}), \\ \geqslant a_1 \mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_{\boldsymbol{\theta}_0}) - a_0 \mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_{\boldsymbol{\theta}}), \end{cases}$$

for some positive numerical constants a_0 and a_1. Note that the mappings $\boldsymbol{\theta} \mapsto a_0 \mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_{\boldsymbol{\theta}_0}) - a_1 \mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_{\boldsymbol{\theta}})$ and $\boldsymbol{\theta} \mapsto a_1 \mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_{\boldsymbol{\theta}_0}) - a_0 \mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_{\boldsymbol{\theta}})$ both reach their maximum at $\overline{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_{\boldsymbol{\theta}})$, i.e. the parameter of the best approximation $\mathbf{Q}_{\overline{\boldsymbol{\theta}}}$ of $\mathbf{Q}^\star$ in our statistical model. Since $\mathbb{E}\left[\mathbf{T}(\boldsymbol{X}, \boldsymbol{\theta}_0, \boldsymbol{\theta})\right]$ is unknown, to build our algorithm, we replace it by its

empirical counterpart $\mathbf{T}(\boldsymbol{X}, \boldsymbol{\theta}_0, \boldsymbol{\theta})$ and define $\boldsymbol{\theta}_1$ as the maximizer of $\boldsymbol{\theta} \mapsto \mathbf{T}(\boldsymbol{X}, \boldsymbol{\theta}_0, \boldsymbol{\theta})$ over $\boldsymbol{\Theta}$. If the quantity

$$\mathbf{T}(\boldsymbol{X}, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1) = \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbf{T}(\boldsymbol{X}, \boldsymbol{\theta}_0, \boldsymbol{\theta}) = \boldsymbol{\upsilon}(\boldsymbol{X}, \boldsymbol{\theta}_0) \leqslant \varepsilon$$

$\boldsymbol{\theta}_0$ is necessarily a ρ-estimator. Otherwise, we iterate the process, taking for $\boldsymbol{\theta}_0$ the value $\boldsymbol{\theta}_1$, and stop as soon as the condition $\boldsymbol{\upsilon}(\boldsymbol{X}, \boldsymbol{\theta}_0) \leqslant \varepsilon$ is satisfied or the number of loops exceeds some given number $L > 0$. In our simulations, we chose $\varepsilon = 1$ and $L = 100$. To find the maximizer of the mapping $\boldsymbol{\theta} \mapsto \mathbf{T}(\boldsymbol{X}, \boldsymbol{\theta}_0, \boldsymbol{\theta})$ we used the CMA (Covariance Matrix Adaptation) method which turned out to be more stable than the gradient descent method. For more details about the CMA method, we refer the reader to Hansen (2016).

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 $\boldsymbol{\upsilon}(\boldsymbol{X}, \widehat{\boldsymbol{\theta}}) > \varepsilon$ 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 $\widehat{\boldsymbol{\theta}}$.

To intialize the process we choose the value of $\boldsymbol{\theta}_0$ as follows. In the case of logit 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)

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

where $e_1, \ldots, e_d$ denotes the canonical basis of $\mathbb{R}^d$. 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 $\boldsymbol{\theta}_0 = \widehat{\boldsymbol{\theta}}_0$.

5.2. **Comparisons of the estimators 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 evaluate the risk

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

of an estimator $\widetilde{\boldsymbol{\theta}}(\boldsymbol{X})$ by the Monte Carlo method. More precisely, we generate a sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{100}$ of size 100 of the distribution $P_{\boldsymbol{\theta}^\star}^{\otimes n}$ of $\boldsymbol{X}$

in order to build 100 i.i.d. copies of the estimator $\widetilde{\boldsymbol{\theta}}(\boldsymbol{X})$. Then we generate independently a sample $W'_1, \ldots, W'_{1000}$ of size 1000 of the distribution P_W and compute

$$\widehat{R}_n(\widetilde{\boldsymbol{\theta}}) = \frac{1}{100}\sum_{i=1}^{100}\left[\frac{1}{1000}\sum_{j=1}^{1000} h^2\left(Q_{\boldsymbol{\theta}^\star}(W'_j), Q_{\widetilde{\boldsymbol{\theta}}(\boldsymbol{X}_i)}(W'_j)\right)\right]. \tag{33}$$

In (33) the Hellinger distances are calculated for all $\theta, \theta' \in I$ according to the formula

$$h^2(Q_\theta, Q_{\theta'}) = 1 - \exp\left[A\left(\frac{\theta+\theta'}{2}\right) - \frac{A(\theta)+A(\theta')}{2}\right] \tag{34}$$

where A is given in (2). For this simulation study $n = 500$.

Logit model. We consider the function $\boldsymbol{\theta}^\star = \boldsymbol{\theta}$ given by (32) with $\eta = (1, \ldots, 1) \in \mathbb{R}^6$ and the distribution $P_W = (P_W^{(1)} + P_W^{(2)} + P_W^{(3)})/3$ 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 (32) with $\eta = (0.7, 3, 4, 10, 2, 5)$ and $P_W = P_{W,1}^{\otimes 2} \otimes P_{W,2} \otimes P_{W,3}^{\otimes 2}$ 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 (32) with $\eta = (0.07, 3, 4, 6, 2, 1)$ and $P_W = P_{W,1}^{\otimes 3} \otimes P_{W,2}^{\otimes 2}$ 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 estimators of interest, we shall first compute the estimated risk $\widehat{R}_n(\widehat{\boldsymbol{\theta}})$ of our estimator and then, given another estimator $\widetilde{\boldsymbol{\theta}}$ (either the MLE or the median-based estimator $\widehat{\boldsymbol{\theta}}_0$), the quantity

$$\mathcal{E}(\widetilde{\boldsymbol{\theta}}) = \frac{\widehat{R}_n(\widetilde{\boldsymbol{\theta}}) - \widehat{R}_n(\widehat{\boldsymbol{\theta}})}{\widehat{R}_n(\widehat{\boldsymbol{\theta}})}. \tag{35}$$

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 small negative values a slight inferiority. The respective values of $\widehat{R}_n(\widehat{\boldsymbol{\theta}})$ and $\mathcal{E}(\widetilde{\boldsymbol{\theta}})$ are displayed in Table 1.

We observe that the quantities $\widehat{R}_n(\widehat{\boldsymbol{\theta}})$ are of order 1.4×10^{-3} in all three cases, which is consistent with the fact that our risk bound (12) only depends on n and the VC-dimension of $\overline{\boldsymbol{\Theta}}$ but not on the particular exponential family when the model is true. The numerical results also indicate that the values of the constants c_j, $j = 1, 2, 3$ in (11) are probably very pessimistic as compared to the "real" ones.

TABLE 1. Values of $\widehat{R}_n(\widehat{\boldsymbol{\theta}})$ and $\mathcal{E}(\widetilde{\boldsymbol{\theta}})$ given by (33) and (35) respectively when the model is well-specified

	$\widehat{R}_n(\widehat{\boldsymbol{\theta}})$	MLE	$\widehat{\boldsymbol{\theta}}_0$
Logit	0.0013	+0.45%	–
Poisson	0.0016	-0.34%	+440%
Exponential	0.0014	-0.58%	+120%

When the model is correct, the risks of the MLE and $\widehat{\boldsymbol{\theta}}$ are quite similar. In fact, a look at the simulations show that the ρ-estimator coincides most of the time with the MLE, a fact which is consistent with the result proved in Baraud *et al* (2017) (Section 5) that states the following: under (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. In this case, both the MLE and the ρ-estimator outperform the estimator $\widehat{\boldsymbol{\theta}}_0$. Moreover, we observe that in all cases the algorithm returns the ρ-estimator after at most two steps.

5.3. **Comparisons of the estimators 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. and simply follow the framework of the previous section with the same distributions $P_{\boldsymbol{\theta}^\star} = Q_{\boldsymbol{\theta}^\star}.P_W$ and parameter values $\boldsymbol{\theta}^\star \in \overline{\boldsymbol{\Theta}}$ we previously used for each of our three exponential families. But we now add to the sample an extra variable $X_n = (W_n, Y_n)$ which is an outlier. We still evaluate the performance of an estimator $\widetilde{\boldsymbol{\theta}}_n(\boldsymbol{X})$ by its empirical risk (33). For the logit regression we take $W_n = 1000(1, 1, 1, 1, 1)$ and $Y_n = 0$, for the Poisson family $W_n = 0.1(1, 1, 1, 1, 1)$ and $Y_n = 200$ and for the exponential distribution $W_n = 5 \times 10^{-3}(1, 1, 1, 10, 10)$ and $Y_n = 1000$. The results are displayed in Table 2. We note that the risks of the ρ-

TABLE 2. Values of $\widehat{R}_n(\widehat{\boldsymbol{\theta}})$ and $\mathcal{E}(\widetilde{\boldsymbol{\theta}})$ given by (33) and (35) respectively in presence of an outlier

	$\widehat{R}_n(\widehat{\boldsymbol{\theta}})$	MLE	$\widehat{\boldsymbol{\theta}}_0$
Logit	0.0014	+15000%	–
Poisson	0.0019	+1900%	+330%
Exponential	0.0015	+7400%	+99%

estimator are quite similar to those given in Table 1 despite the presence of an outlier among the data set. The performance of $\widehat{\boldsymbol{\theta}}$ remains much better than that of $\widehat{\boldsymbol{\theta}}_0$. As expected, the MLE behaves poorly.

TABLE 3. Quartiles for the number of loops 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 3 displays the quartiles of the empirical distributions, based on our 100 simulations, of the number of loops l that have been necessary for our algorithm to compute the ρ-estimator. It shows that the presence of the outlier has slightly increased the number of loops that have been necessary to compute the ρ-estimator. We recall that when the model was exact, at most two loops were necessary.

5.4. **Comparisons of the estimators when the data are contaminated.** We now set $n = 500$ and keep the respective values of the distributions $P_{\boldsymbol{\theta}^\star}$ to those of Section 5.2, but we now assume that the true distribution of the i.i.d. random variables $X_1, \ldots, X_n$ is $P^\star = 0.95P_{\boldsymbol{\theta}^\star} + 0.05R$ for some distribution R on $\mathscr{W} \times \mathscr{Y}$, thus allowing an amount of contamination of 5%. In such a situation, the squared Hellinger distance between the true distribution $P^\star$ and the model is bounded by $h^2(P^\star, P_{\boldsymbol{\theta}^\star}) \leqslant 5\%$ according to (18). We assume that R is the distribution of a pair of random variables (W', Y'), W' with distribution P_W and the conditional distribution of Y' given $W' = w$ admitting a density r_w with respect to ν. As before, we evaluate the risk

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

of an estimator $\widetilde{\boldsymbol{\theta}}(\boldsymbol{X})$ by the Monte Carlo method. To compute the integral involved in the definition of the Hellinger distance

$$h^2\left(Q^\star(w), Q_{\widetilde{\boldsymbol{\theta}}(\boldsymbol{X})}(w)\right) = \frac{1}{2}\int_{\mathscr{Y}} \left(\sqrt{0.95q_{\boldsymbol{\theta}^\star(w)}(y) + 0.05r_w(y)} - \sqrt{q_{\widetilde{\boldsymbol{\theta}}(w)}(y)}\right)^2 d\nu(y)$$

we used a numerical approximation. For each estimator $\widetilde{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}^\star$, we obtain an estimation $\widetilde{R}_n(\widetilde{\boldsymbol{\theta}})$ of $R_n(\widetilde{\boldsymbol{\theta}})$. As before, in order to compare the performance of the other estimators to $\widehat{\boldsymbol{\theta}}$, we evaluate

$$\mathcal{E}'(\widetilde{\boldsymbol{\theta}}) = \frac{\widetilde{R}_n(\widetilde{\boldsymbol{\theta}}) - \widetilde{R}_n(\widehat{\boldsymbol{\theta}})}{\widetilde{R}_n(\widehat{\boldsymbol{\theta}})}. \tag{36}$$

In the Poisson case, the random variable Y' writes as $80 + B$ 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, Y' is independent of W and uniformly distributed on $[50, 60]$.

With these values of the distribution R, the approximation error of the model, which is bounded by $h^2(P^\star, P_{\boldsymbol{\theta}^\star})$ is not larger than 0.025 in the Poisson case and 0.029 for the exponential distribution, hence smaller than 5% as expected. The corresponding results are presented in Table 4.

TABLE 4. Values of $\widetilde{R}_n(\widehat{\boldsymbol{\theta}})$ and $\mathcal{E}'(\widetilde{\boldsymbol{\theta}})$ given by (36) when (in average) 5% of the data are contaminated

	$\widetilde{R}_n(\widehat{\boldsymbol{\theta}})$	MLE	$\widehat{\boldsymbol{\theta}}_0$
Poisson	0.027	+760%	+11%
Exponential	0.039	+340%	-15%

For both models we see that the the risk of the ρ-estimator is of the same order of magnitude as that of the approximation term $h^2(P^\star, \overline{\mathscr{P}}) \leqslant h^2(P^\star, P_{\boldsymbol{\theta}^\star})$. As before, the MLE behaves poorly. Its risk explodes with the contamination while the performance of the median based estimator $\widehat{\boldsymbol{\theta}}_0$ is comparable to that of the ρ-estimator.

In Table 5, we observe that the number of loops for calculating the ρ-estimator increases substantially as compared to the two previous situations. Our algorithm seems to have more difficulties to reach the ρ-estimator. Nevertheless, the fact that it sometimes has to stop when $l = 100$, i.e. before reaching its goal, does not alter its final performance as described by Table 4.

TABLE 5. Quartiles for the number of loops when the data are contaminated

	1st Quartile	Median	3rd Quartile	Maximum
Poisson	5	5	5	100
Exponential	5	11	40	100

In conclusion, in our examples, we have seen that if the model is exact the MLE performs well but cannot deal with a slight misspecification of it. The presence of a single outlier explodes its risk. The median-based estimator $\widehat{\boldsymbol{\theta}}_0$ performs poorly as compared to the ρ-estimator when the model is exact and when the data set 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 $\widehat{\boldsymbol{\theta}}_0$, only the ρ-estimator shows some good and stable estimation properties in the situations we have studied.

6. An upper bound on the expectation of the supremum of an empirical process over a VC-subgraph class

The aim of this section is to prove the following result.

Theorem 2. *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 $V \geqslant 1$. If*

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

then

$$\mathbb{E}\left[Z(\mathscr{F})\right] \leqslant 4.74\sqrt{nV\sigma^2\mathscr{L}(\sigma)} + 90V\mathscr{L}(\sigma), \tag{37}$$

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

Let us now turn to the proof. It follows from classical symmetrisation arguments that $\mathbb{E}\left[Z(\mathscr{F})\right] \leqslant 2\mathbb{E}\left[\overline{Z}(\mathscr{F})\right]$, 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

$$\mathbb{E}\left[\overline{Z}(\mathscr{F})\right] \leqslant 2.37\sqrt{nV\sigma^2\mathscr{L}(\sigma)} + 45V\mathscr{L}(\sigma). \tag{38}$$

Given a probability P and a class of functions $\mathscr{G}$ on a $(E, \mathcal{E})$ 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 $g \in \mathscr{G}$

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

We start with the following lemma.

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

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

Proof of Lemma 1. 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 $x \in \mathscr{X}$

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

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 $[-1,1]$,

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

and consequently, for all $\varepsilon > 0$

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

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 is based on a chaining argument. It follows from the monotone convergence theorem that it is actually enough to prove (38) with $\mathscr{F}_J$, $J \geqslant 1$, in place of $\mathscr{F}$ where $(\mathscr{F}_J)_{J\geqslant 1}$ is a sequence of finite subsets of $\mathscr{F}$ which is increasing for the inclusion and satisfies $\bigcup_{J\geqslant 1} \mathscr{F}_J = \mathscr{F}$. 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 $\boldsymbol{X} = (X_1, \ldots, X_n)$. We set

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

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 the $\mathbb{L}_2(\mathscr{X}, \mathcal{X}, P_{\boldsymbol{X}})$-norm denoted $\|\cdot\|_{2,\boldsymbol{X}}$. In particular, we can associate to a function $f \in \mathscr{F}$ a sequence $(f_k)_{k\geqslant 1}$ with $f_k \in \mathscr{F}_k$ satisfying $\|f - f_k\|_{2,\boldsymbol{X}} \leqslant q^k\widehat{\sigma}$ for all $k \geqslant 1$. Actually, since $\mathscr{F}$ is finite $f_k = f$ for all k large enough. Besides, it follows from Lemma 1 with the choices $r = 2$ and $P = P_{\boldsymbol{X}}$ that for all $k \geqslant 1$ 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

$$h(\epsilon) = \log\left[e(V+1)(2e)^V\right] + 2V\log\left(\frac{2}{\epsilon}\right) \quad \text{for all } \epsilon \in (0,1]. \tag{39}$$

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 $k \geqslant 1$, we deduce that

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

and consequently,

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

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

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

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

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

we deduce that

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

and for all $k \geqslant 1$

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

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

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

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

$$\begin{aligned}
\mathbb{E}\left[\overline{Z}(\mathscr{F})\right] &\leqslant \sqrt{2n}\frac{1+q}{1-q}\mathbb{E}\left[G(\widehat{\sigma})\right] \leqslant \sqrt{2n}\frac{1+q}{1-q}G\left(\mathbb{E}\left[\widehat{\sigma}\right]\right) \\
&\leqslant \sqrt{2n}\frac{1+q}{1-q}G\left(\sqrt{\mathbb{E}\left[\widehat{\sigma}^2\right]}\right). \qquad (40)
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}\left[n\widehat{\sigma}^2\right] &\leqslant \mathbb{E}\left[\sup_{f\in\mathscr{F}}\sum_{i=1}^n\left(f^2(X_i) - \mathbb{E}\left[f^2(X_i)\right]\right)\right] + \sup_{f\in\mathscr{F}}\sum_{i=1}^n \mathbb{E}\left[f^2(X_i)\right] \\
&\leqslant 2\mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \varepsilon_i f^2(X_i)\right|\right] + n\sigma^2 \\
&\leqslant 8\mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right] + n\sigma^2 = 8\mathbb{E}\left[\overline{Z}(\mathscr{F})\right] + n\sigma^2 \qquad (41)
\end{aligned}$$

and we infer from (40) that

$$\mathbb{E}\left[\overline{Z}(\mathscr{F})\right] \leqslant \sqrt{2n}\frac{1+q}{1-q}G\left(B\right) \quad \text{with} \quad B = \sqrt{\sigma^2 + \frac{8\mathbb{E}\left[\overline{Z}(\mathscr{F})\right]}{n}} \wedge 1. \qquad (42)$$

The following lemma provides an evaluation of G.

Lemma 2. *Let a, b, y_0 be positive numbers and $y \in [y_0, 1]$,*

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

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

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

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

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 with $y_0 = \sigma$ and $y = B$ and deduce from (42) that

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

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

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

we get that

$$\mathbb{E}\left[\overline{Z}(\mathscr{F})\right] \leqslant 8A^2 + \sqrt{64A^4+2A^2n\sigma^2} \leqslant 16A^2 + A\sqrt{2n\sigma^2}. \tag{43}$$

Finally, we conclude by using the inequalities

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

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

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

and together with (43) lead to (38).

7. Proofs

7.1. **Proof of Theorem 1.** We recall that the function ψ defined by (6) 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 with Proposition 3). Theorem 1 is 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 $\boldsymbol{\mathcal{P}}$ the following families of densities (with respect to $\boldsymbol{\mu}$) on $\mathscr{X}^n = (\mathscr{W} \times \mathscr{Y})^n$

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

and by $\mathscr{P}$ the corresponding ρ-model with representation $(\boldsymbol{\mu}, \boldsymbol{\mathcal{P}})$.

Let us first prove

Proposition 5. *Under Assumption 1, the class of functions* $\mathcal{P} = \{q_{\boldsymbol{\theta}} : (w, y) \mapsto q_{\boldsymbol{\theta}(w)}(y),\ \boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}}\}$ *on* $\mathscr{X} = \mathscr{W} \times \mathscr{Y}$ *is VC-subgraph with dimension not larger than* $9.41V$.

Proof. The exponential function being monotone, 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 map on I defined 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

$$C_{\boldsymbol{\theta}}^+ = \{(w, b, b') \in \mathscr{W} \times \overline{I}^2,\ \boldsymbol{\theta}(w) > b\}$$
$$C_{\boldsymbol{\theta}}^- = \{(w, b, b') \in \mathscr{W} \times \overline{I}^2,\ \boldsymbol{\theta}(w) < b'\}$$

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 $i \in \{1, \dots, k\}$, necessarily the b_i belong to $\mathbb{R}$ for all $i \in \{1, \dots, k\}$. 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 $\widetilde{z}_1 = (w_1, b_1), \dots, \widetilde{z}_k = (w_k, b_k)$ in $\mathscr{W} \times \mathbb{R}$. This is possible only for $k \leqslant V$ since, by Assumption 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 van der Vaart and Wellner (2009) Theorem 1.1 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 smaller than $9.41V$.

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

$$\begin{aligned} j \in J \quad &\Longleftrightarrow \quad S(y_j)\boldsymbol{\theta}(w_j) - A\left(\boldsymbol{\theta}(w_j)\right) > u_j \iff \boldsymbol{\theta}(w_j) \in (\underline{a}(y_j, u_j), \overline{a}(y_j, u_j)) \\ &\Longleftrightarrow \quad z_j = (w_j, \underline{a}(y_j, u_j), \overline{a}(y_j, u_j)) \in C^+_{\boldsymbol{\theta}} \cap C^-_{\boldsymbol{\theta}}\,. \end{aligned}$$

This means that 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 $\{z_1, \dots, z_k\}$ in $\mathscr{Z}$. This is possible for $k \leqslant 9.41V$ 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 of $\mathscr{P}$.

Proposition 6. *Under Assumption 1, for all product probabilities* $\mathbf{P}, \overline{\mathbf{P}} = \otimes_{i=1}^n \overline{P}_i$ *on* $(\mathscr{X}^n, \mathcal{X}^{\otimes n})$ *with* $\overline{P}_i = \overline{p} \cdot \mu_i$ *for all* $i \in \{1, \dots, n\}$,

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

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 $y > 0$,

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

It follows from Proposition 5 and Baraud *et al* (2017)[Proposition 42] 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 $\sup_{f\in\mathscr{F}_y} n^{-1}\sum_{i=1}^n \mathbb{E}\left[f(X_i)\right] \leqslant \sigma^2(y) = (a_2^2y^2/n)\wedge 1$. Applying Theorem 2 with $\mathscr{F}=\mathscr{F}_y$, we obtain that

$$\begin{aligned} w^{\mathscr{P}}(\mathbf{P},\overline{\mathbf{P}},y) &= \mathbb{E}\left[\sup_{f\in\mathcal{F}_y}\left|\sum_{i=1}^n f(X_i)-\mathbb{E}\left[f(X_i)\right]\right|\right] \\ &\leqslant 4.74a_2y\sqrt{\overline{V}\mathscr{L}(\sigma(y))}+90\overline{V}\mathscr{L}(\sigma(y)) \\ &= 14.55a_2y\sqrt{V\mathscr{L}(\sigma(y))}+846.9V\mathscr{L}(\sigma(y)). \end{aligned}$$

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

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

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

$$\begin{aligned} w^{\mathscr{P}}(\mathbf{P},\overline{\mathbf{P}},y) &\leqslant 14.55a_2y\sqrt{VL}+276.1VL \\ &= \frac{a_1y^2}{8}\left[\frac{8\times 14.55a_2\sqrt{VL}}{a_1y}+\frac{8\times 846.9VL}{a_1y^2}\right] \\ &\leqslant \frac{a_1y^2}{8}\left[\frac{8\times 14.55a_2\sqrt{VL}}{a_1\beta^{-1}\sqrt{D}}+\frac{8\times 846.9VL}{a_1\beta^{-2}D}\right] \\ &= \frac{a_1y^2}{8}\left[\frac{2\times 14.55\sqrt{VL}}{\sqrt{D}}+\frac{8\times 846.9a_1VL}{16a_2^2D}\right] \\ &= \frac{a_1y^2}{8}\left[\frac{29.1\sqrt{VL}}{\sqrt{D}}+\frac{37.5VL}{D}\right]\leqslant\frac{a_1y^2}{8} \end{aligned}$$

for $D=10^3VL>2^{-11}V$. The result follows from the definition of the ρ-dimension in Baraud and Birgé (2018)[Definition 4]. □

Let us now end the proof of Theorem 1. It follows from Baraud and Birgé (2018)[Theorem 1], 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 3.1, satisfies for all $\overline{\mathbf{P}}\in\mathscr{P}$, with a probability at least $1-e^{-\xi}$,

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

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 $D^{\mathscr{P}}(\mathbf{P}^\star, \overline{\mathbf{P}}) \leqslant 10^3 V \left[9.11 + \log_+(n/V)\right]$ by Proposition 6. Finally, the result follows from the fact that $\mathbf{h}^2(\mathbf{P}^\star, \widehat{\mathbf{P}}) = \mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_{\widehat{\boldsymbol{\theta}}})$ and $\mathbf{h}^2\left(\mathbf{P}^\star, \mathscr{P}\right) = \mathbf{h}^2\left(\mathbf{Q}^\star, \mathscr{Q}\right)$.

7.2. **A preliminary result.** The following result holds.

Proposition 7. *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*

$$G_\varepsilon(x) = L \sum_{k=0}^{2^N-1} \varepsilon_{k+1} g\left(2^N x - k\right) \quad \text{for all } x \in [0,1]. \tag{44}$$

Then, G_ε satisfies (24) with $\alpha \in (0,1]$ and $M > 0$ provided that $L \leqslant 2^{-[(N-1)\alpha+1]} M$.

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 $0 \leqslant y - x \leqslant 2^{-N}$ and the fact that $L 2^{N\alpha} \leqslant L 2^{(N-1)\alpha+1} \leqslant M$, we obtain that

$$\begin{aligned} |G_\varepsilon(y) - G_\varepsilon(x)| &= L\,|g_k(y) - g_k(x)| \leqslant L 2^N (y-x) \\ &\leqslant L 2^N (y-x)^{1-\alpha} (y-x)^\alpha \leqslant L 2^{N\alpha} (y-x)^\alpha \leqslant M (y-x)^\alpha. \end{aligned}$$

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

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

and since g vanishes at 0 and 1,

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

and the conclusion follows since $L \leqslant 2^{-[(N-1)\alpha+1]} M$. □

We shall often use the following version of Assouad's lemma.

Lemma 3 (Assouad's Lemma)**.** *Let $\mathcal{P}$ be a family of probabilities on a measurable space $(\mathscr{X}, \mathcal{X})$. Assume that for some integer $d \geqslant 1$, $\mathcal{P}$ 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 $\varepsilon, \varepsilon' \in \{-1,1\}^d$

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

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

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

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

$$\sup_{P\in\mathcal{P}} \mathbb{E}_{\mathbf{P}}\left[h^2(P,\widehat{P}(\boldsymbol{X}))\right] \geqslant \frac{d\eta}{8}\max\left\{1-\sqrt{2a}, (1-a/n)^{2n}\right\}, \tag{45}$$

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 $\{-1,1\}^d$ the mapping $\varepsilon \mapsto h^2(P,P_\varepsilon)$. By definition of $\overline{\varepsilon}$,

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

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

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

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 all $i \in \{1,\ldots,d\}$. 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 $\varepsilon \in \{0,1\}$. □

7.3. **Proof of Proposition 1.** Using that for all $x \in [0,1]$, $(1-e^{-1})x \leqslant 1-e^{-x} \leqslant x$ for all $x \in [0,1]$, we deduce from (25) that

$$\frac{1}{2}(1-e^{-1})\left\|\sqrt{\boldsymbol{\gamma}} - \sqrt{\boldsymbol{\gamma}'}\right\|_2^2 \leqslant h^2(R_\gamma, R_{\gamma'}) \leqslant \frac{1}{2}\left\|\sqrt{\boldsymbol{\gamma}} - \sqrt{\boldsymbol{\gamma}'}\right\|_2^2. \tag{46}$$

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

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 (44) 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 7, $\boldsymbol{\gamma}_\varepsilon$ satisfies (24) provided that $L \leqslant 2^{-[(N-1)\alpha+1]}M$. Hence, under the conditions $L \leqslant 2^{-[(N-1)\alpha+1]}M$ and $b<1$, $\boldsymbol{\gamma}_\varepsilon$ belongs to $\mathcal{H}_\alpha(M)$ for all $\varepsilon \in \{-1,1\}^{|\Lambda|}$. For all $\varepsilon,\varepsilon' \in \{-1,1\}^{|\Lambda|}$,

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

and

$$\left|\sqrt{\gamma_\varepsilon}-\sqrt{\gamma_{\varepsilon'}}\right| \leqslant \sqrt{(1+b)L}-\sqrt{(1-b)L} = \left[\sqrt{1+b}-\sqrt{1-b}\right]\sqrt{L}.$$

In particular, $\left\|\sqrt{\gamma_\varepsilon}-\sqrt{\gamma_{\varepsilon'}}\right\|_\infty \leqslant 1$ for

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

and writing R_ε for R_{γ_ε} for short, it follows from (46) that

$$\frac{(1-e^{-1})}{8(1+b)L}\left\|G_\varepsilon - G_{\varepsilon'}\right\|_2^2 \leqslant h^2(R_\varepsilon, R_{\varepsilon'}) \leqslant \frac{1}{8(1-b)L}\left\|G_\varepsilon - G_{\varepsilon'}\right\|_2^2. \tag{47}$$

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 $\varepsilon, \varepsilon' \in \{-1,1\}^{|\Lambda|}$

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

Let us denote by $P_\gamma = R_\gamma \cdot P_W$ the probability associated to R_γ and write P_ε for P_{γ_ε} for short. We deduce from (47) that provided that L and b satisfies

$$L \leqslant \left(2^{-[(N-1)\alpha+1]}M\right) \wedge \frac{1+\sqrt{1-b^2}}{2b^2} \wedge \frac{(1-b)2^{N-3}}{\left\|g\right\|_2^2 n} \tag{48}$$

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

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

We derive from the equalities

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

and (45) that

$$\mathcal{R}_n(\mathcal{H}_\alpha(M)) \geqslant \frac{(1-e^{-1})\left\|g\right\|_2^2 L}{16(1+b)}\left(1-\sqrt{2a}\right) \geqslant \frac{(1-e^{-1})\left\|g\right\|_2^2 L}{32(1+b)}. \tag{49}$$

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

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

and $N \geqslant 2$. Otherwise, we choose $N = 1$. Note that in any case,

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

Besides, if $N \geqslant 2$

$$2^{-[(N-1)\alpha+1]}M = 2^{-1}M2^{-(N-1)\alpha} \geqslant 2^{-1}M\left[\frac{2^{2+\alpha}\|g\|_2^2 Mn}{1-b}\right]^{-\alpha/(1+\alpha)}$$
$$= \left(\frac{M}{2}\right)^{1/(1+\alpha)}\left[\frac{1-b}{2^{3+\alpha}\|g\|_2^2 n}\right]^{\frac{\alpha}{1+\alpha}} = L_1$$

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 (48), and we derive from (49) that

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

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 $\|g\|_2^2 = 1/12$.

7.4. **Proof of Proposition 2.** Since the statistical model $\overline{\mathscr{Q}} = \{R_\gamma = Q_{u(\gamma)},\ \gamma \in J\}$ is regular with constant Fisher information equal to 8, by applying Theorem 7.6 page 81 in Ibragimov and Has'minskiĭ (1981) we obtain that

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

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

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

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

7.5. **Proof of Proposition 3.** Let $\overline{\boldsymbol{\Gamma}} = \overline{\boldsymbol{\Gamma}}_D$ be the linear space of functions which are piecewise constant on each element of a partition $\{I_j, j \in \{1,\ldots,D\}\}$ of $[0,1]$ into $D \geqslant 1$ intervals of lengths $1/D$. The value of D will be chosen later on. Let $\boldsymbol{\Gamma}$ be a countable and dense subset of $\overline{\boldsymbol{\Gamma}}$ with respect to the supremum norm $\|\cdot\|_\infty$, i.e. $\|\boldsymbol{\gamma}\|_\infty = \sup_{w\in\mathscr{W}}|\boldsymbol{\gamma}(w)|$ for all functions $\boldsymbol{\gamma}$ on $\mathscr{W} = [0,1]$. 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 $\overline{\boldsymbol{\gamma}} = \sum_{j=1}^D \gamma_j \mathbb{1}_{I_j} \in \overline{\boldsymbol{\Gamma}}$. Since for all $w \in I_j$, $|\boldsymbol{\gamma}(w) - \overline{\boldsymbol{\gamma}}(w)| \leqslant \sup_{|w-w'|\leqslant 1/D}|\boldsymbol{\gamma}(w) - \boldsymbol{\gamma}(w')| \leqslant MD^{-\alpha}$ and $\boldsymbol{\Gamma}$ is dense in $\overline{\boldsymbol{\Gamma}}$

$$\sup_{\boldsymbol{\gamma}\in\mathcal{H}_\alpha(M)}\inf_{\overline{\boldsymbol{\gamma}}\in\overline{\boldsymbol{\Gamma}}}\|\boldsymbol{\gamma} - \overline{\boldsymbol{\gamma}}\|_2 \leqslant \sup_{\boldsymbol{\gamma}\in\mathcal{H}_\alpha(M)}\inf_{\overline{\boldsymbol{\gamma}}\in\overline{\boldsymbol{\Gamma}}}\|\boldsymbol{\gamma} - \overline{\boldsymbol{\gamma}}\|_\infty$$
$$= \sup_{\boldsymbol{\gamma}\in\mathcal{H}_\alpha(M)}\inf_{\overline{\boldsymbol{\gamma}}\in\boldsymbol{\Gamma}}\|\boldsymbol{\gamma} - \overline{\boldsymbol{\gamma}}\|_\infty \leqslant MD^{-\alpha}.$$

Using (29) 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,

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

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

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

Let us set $L_n = \log(en)$ and choose $D \geqslant 1$ such that

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

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

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

7.6. **Proof of Proposition 4.** Let a_0 be the middle of the interval K of length $2\overline{L}$. Given $N \geqslant 1$ and $L > 0$, we define $\boldsymbol{\gamma}_\varepsilon = a_0 + G_\varepsilon$ where G_ε is defined in Proposition 7 for all $\varepsilon \in \{-1,1\}^{2^N}$. Provided that $L \leqslant \overline{L} \wedge L_0$ with $L_0 = 2^{-[(N-1)\alpha+1]}M$, the functions $\boldsymbol{\gamma}_\varepsilon$ takes their values in $K \subset J$ and satisfies (24) and consequently belongs to $\mathcal{H}_\alpha(M)$ for all $\varepsilon \in \{-1,1\}^{2^N}$. Let $R_\varepsilon = R_{\boldsymbol{\gamma}_\varepsilon}$ for all $\varepsilon \in \{-1,1\}^{2^N}$ and, as in the proof of Proposition 1, we set $P_{\boldsymbol{\gamma}} = R_{\boldsymbol{\gamma}} \cdot P_W$ and $P_{\boldsymbol{\gamma}_\varepsilon} = P_{\boldsymbol{\gamma}_\varepsilon}$ for short. Integrating the inequalities (29) and (30)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

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

Since P_W is the uniform distribution on $[0,1]$, by arguing as in the proof of Proposition 1

$$\|G_\varepsilon - G_{\varepsilon'}\|_2^2 = 4L^2 2^{-N} \|g\|_2^2 \delta(\varepsilon, \varepsilon') \quad \text{for all } \varepsilon, \varepsilon' \in \{-1,1\}^{2^N}$$

and consequently, provided that L satisfies

$$L \leqslant \overline{L} \wedge L_0 \wedge \left(4\kappa \|g\|_2 \sqrt{2^{-(N-1)} n}\right)^{-1} \tag{50}$$

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 Assouad's lemma with $d = 2^N$,

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

We derive from (45) that

$$\mathcal{R}_n(\mathcal{H}_\alpha(M)) \geqslant \frac{c_K^2\|g\|_2^2 L^2}{2}\left(1-\sqrt{2a}\right) \geqslant \frac{c_K^2\|g\|_2^2 L^2}{4}. \tag{51}$$

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

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

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

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

When $N \geqslant 2$,

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

while $L_0 = M/2$ when $N=1$. The choice $L = \overline{L} \wedge L_1 \wedge (M/2)$ satisfies (50) and we deduce from the equalities

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

and (51) that

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

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

References

Antoniadis, A. and Sapatinas, T. (2001). Wavelet shrinkage for natural exponential families with quadratic variance functions. *Biometrika*, 88(3):805–820.

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

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

Birgé, L. (1986). On estimating a density using Hellinger distance and some other strange facts. *Probab. Theory Relat. Fields*, 71(2):271–291.

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

Candès, E. and Sur, P. (2020). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics*, 48(1):27–42.

Hansen, N. (2016). *The CMA Evolution Strategy: A Tutorial*.

Haussler, D. (1995). Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *J. Combin. Theory Ser. A*, 69(2):217–232.

Ibragimov, I. A. and Has'minskiĭ, R. Z. (1981). *Statistical Estimation. Asymptotic Theory*, volume 16. Springer-Verlag, New York.

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

Kroll, M. (2019). Non-parametric Poisson regression from independent and weakly dependent observations by model selection. *J. Statist. Plann. Inference*, 199:249–270.

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.

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.

van der Vaart, A. 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.*, pages 103–107. Inst. Math. Statist., Beachwood, OH.

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.

Research Unit in Mathematics,
University of Luxembourg
Maison du nombre
6 avenue de la Fonte
L-4364 Esch-sur-Alzette
Grand Duchy of Luxembourg
Email address: yannick.baraud@uni.lu